Received: 7 August 2019 | Revised: 16 November 2020 | Accepted: 31 January 2021

DOI: 10.1002/sim.8915

RESEARCH ARTICLE

Statistics in Medicine WILEY

Selecting the model for multiple imputation of missing data: Just use an IC!

Firouzeh Noghrehchi[1] | **Jakub Stoklosa**[2,3] | **Spiridon Penev**[2] | **David I. Warton**[2,3]

[1]Discipline of Biomedical Informatics and Digital Health, The University of Sydney, Sydney, New South Wales, Australia

[2]School of Mathematics and Statistics, The University of New South Wales, Sydney, New South Wales, Australia

[3]Evolution and Ecology Research Centre, The University of New South Wales, Sydney, New South Wales, Australia

Correspondence
Firouzeh Noghrehchi, Discipline of Biomedical Informatics and Digital Health, The University of Sydney, NSW, Australia.
Email: firouzeh.noghrehchi@sydney.edu.au

Funding information
Australian Government Research Training Program Scholarship

Multiple imputation and maximum likelihood estimation (via the expectation-maximization algorithm) are two well-known methods readily used for analyzing data with missing values. While these two methods are often considered as being distinct from one another, multiple imputation (when using improper imputation) is actually equivalent to a stochastic expectation-maximization approximation to the likelihood. In this article, we exploit this key result to show that familiar likelihood-based approaches to model selection, such as Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), can be used to choose the imputation model that best fits the observed data. Poor choice of imputation model is known to bias inference, and while sensitivity analysis has often been used to explore the implications of different imputation models, we show that the data can be used to choose an appropriate imputation model via conventional model selection tools. We show that BIC can be consistent for selecting the correct imputation model in the presence of missing data. We verify these results empirically through simulation studies, and demonstrate their practicality on two classical missing data examples. An interesting result we saw in simulations was that not only can parameter estimates be biased by misspecifying the imputation model, but also by *overfitting* the imputation model. This emphasizes the importance of using model selection not just to choose the appropriate type of imputation model, but also to decide on the appropriate level of imputation model complexity.

KEYWORDS
imputation model selection, information criteria, missing data analysis, stochastic EM algorithm

1 | INTRODUCTION

Missing data commonly arise in many health research datasets, for example, in clinical trials,[1,2] in biomarker data,[3,4] in covariates when fitting Cox models,[5,6] or in longitudinal studies.[7] Multiple imputation (MI)[8] and maximum likelihood estimation (MLE) via the expectation-maximization (EM) algorithm[9] are two well-known methods used to account for missing values in partially observed datasets. MI is a Monte Carlo approach where missing values are imputed to create

a complete dataset, whereas the EM algorithm is an iterative method developed for finding the MLE in the presence of missing data. Stochastic versions of EM, in particular Monte Carlo EM (MCEM)[10] and stochastic EM (StEM)[11-13] are numerical approaches of EM to compute a Monte Carlo approximation to MLE.

A little-appreciated idea is that MI and MLE are closely connected, with some exceptions in comparing their performances in numerical results, for example, see Honaker and King.[14] Specifically, StEM can be understood as a type of multiple imputation where data are imputed from a model that fixes parameter estimates at the maximizer of the imputation model likelihood (a type of "improper MI," see Rubin[8]). This result has been referred to in Reference 15 and 16, and importantly, it permits us to think of the multiple imputation procedure in a likelihood framework, motivating the application of standard likelihood machinery in multiple imputation. While this opportunity offers considerable potential, little work along these lines has been done to date. In particular, the literature on imputation model selection within the MI framework is surprisingly sparse given its potential application range. With the exception of some pragmatic suggestions for model selection,[17-19] no overall and general valid framework has been provided for either model selection or for imputation model selection when using MI. In this article, we focus on the question of selecting an imputation model to use to impute missing values with MI, and show that standard likelihood-based model selection tools are applicable to this problem. We leave the development of a valid framework for model selection for predictors with MI for future work.

When using MI, careful consideration of which imputation model to select is needed, because using the wrong imputation model can result in incorrect inferences and misleading conclusions.[19-21] There are guidelines for specification of the imputation model that tend to be ad hoc and based on heuristic arguments[22-24] rather than providing an overall valid theoretical framework for imputation model selection. Other methods include assessing the sensitivity of MI results using a weighting approach[25] or the pattern-mixture approach,[26,27] and using variable selection procedures to select the best imputation model based on the fraction of missing information when estimated from an appropriate Proxy pattern-mixture model.[28] A key focus of the current study is to provide a flexible, data-driven approach for choosing the imputation model.

In this article, we exploit the equivalence between MI and MLE, to propose imputation model selection using likelihood-based information criteria. In principle, any likelihood-based information criterion with desirable properties could be used, and we focus in particular on the Akaike information criterion (AIC) and Bayesian information criterion (BIC), computed using the observed data likelihood. We show that BIC preserves its usual property of consistency in model selection, when selecting an imputation model, and illustrate this via simulation.

In Sections 2.1 and 2.2, we give an overview of the connection between MI and StEM methods. In Section 3, we develop AIC and BIC for choosing the best imputation model and study some of their properties. Three simulation studies are given in Section 4 and we present some real-data examples in health research in Section 5. Finally, we provide a summary discussion in Section 6.

2 | MISSING DATA ANALYSIS METHODS

Suppose y is the observed data, z is the missing data, and r is the missingness indicator vector where its components are set as $r_i = 0$ if the ith data point is missing and $r_i = 1$ if the ith data point is observed. Note that y can contain the vector of observed response variables, denoted Y, as well as the vector of observed predictors $X_1, X_2, \ldots$. Also, denote θ as a vector of unknown model parameters. For now we assume that these data are missing not at random (MNAR) but note that the methods presented below are also applicable for missing at random (MAR) data, the special case of MNAR where missingness is ignorable. These data are considered to be MAR when $\forall \theta, r, y, z, z^*$ we have $p(r|y, z, \theta) = p(r|y, z^*, \theta)$, and MNAR when $\exists \theta, r, y, z \neq z^*$ such that we have $p(r|y, z, \theta) \neq p(r|y, z^*, \theta)$. For more details on the definitions of MAR and MNAR, see Seaman et al,[29] Mealli et al,[30] and Doretti et al.[31]

To estimate θ, we maximize the observed data likelihood, $p(y, r|\theta)$, which, in the presence of missing data, is obtained by integrating out the missing data from the complete-data likelihood, $p(y, z, r|\theta)$:

$$L(\theta; y, r) = p(y, r|\theta) = \int p(y, z, r|\theta) \mathrm{d}z. \tag{1}$$

When data are MAR and missingness is ignorable, then the missingness mechanism does not depend on the missing data, that is, $p(r|y, z, \theta) = p(r|y, \theta)$. Therefore, the observed likelihood ignoring the missingness mechanism is defined to be $L(\theta; y, r) \propto L(\theta; y) = p(y|\theta) = \int p(y, z|\theta) \mathrm{d}z$, which does not involve the model for r.[27]

Likelihood (1) can be difficult to solve for complex models or when the data are of high dimension. In the missing data context, a common approach is to impute (fill-in) missing data, z, with some plausible values that are a good summary of z, to which we obtain a (pseudo-) complete dataset. Two well-known methods that use imputation procedures are multiple imputation (MI) and stochastic EM (StEM). These two methods turn out to have a close resemblance, although they were developed to serve different purposes. MI was primarily developed in the context of missing data analysis to reflect uncertainty in the imputed values.[8] Stochastic EM was developed more broadly as a computational device when using an EM algorithm. Specifically, StEM was designed to overcome computational difficulties, or intractable E-steps.[10,12,32,33] The objective of both methods is to achieve valid statistical inferences, rather than optimal predictions of the missing data.[34] First, we give a brief overview of both methods, then show their equivalence.

2.1 | Multiple imputation

Multiple imputation is a Monte Carlo method originally proposed by Rubin[8,35] where every missing observation in the dataset is imputed with a simulated value to create a complete dataset. The imputation step is repeated $M \geq 2$ times, such that M-completed datasets are generated. That is, we impute z with a set of plausible values, $z^* = (z^{(1)}, z^{(2)}, \ldots, z^{(M)})$, for some $M \geq 2$. Each (pseudo-) complete dataset is then separately analyzed by standard complete-data analysis methods. In order to achieve asymptotically valid statistical inference, the resulting estimates from each M-completed dataset need to be pooled according to Rubin's rule,[8] see Appendix B for further details.

MI is sometimes performed in an iterative manner to approximate the observed posterior of θ, as summarized in Box 1 (left). Consider a complete data model $p(y, z, r|\theta)$, and the imputation model $p(z|y, r, \theta)$. Multiple imputations are repeated random draws from the posterior predictive distribution of missing data given the observed data and the current estimates of θ. The Monte Carlo average of the current multiple imputed datasets gives a current approximation to the observed posterior, $p(\theta|y, r) = \int p(\theta|y, z, r)p(z|y, r)\mathrm{d}z \simeq M^{-1}\sum_{j=1}^{M} p(\theta|y, z^{(j)}, r)$. Since we have $p(z|y, r) = \int p(z|y, r, \theta)p(\theta|y, r)\mathrm{d}\theta$, these two steps of imputation (I-step) and posterior re-estimation (P-step) are iterated a large number of times to eventually produce a draw of (z, θ) from their joint observed posterior. Finally, the next M imputations are implemented as multiple imputations in Rubin's combination rules to make inference about the dataset.

The algorithm in Box 1 (left) imputes missing values from an imputation model whose parameters were sampled from the posterior distribution for θ. This is referred to by Rubin[8] as "proper MI." Other methods of updating θ are possible, but unless they satisfy Rubin's rules[8] they are labeled "improper" and their statistical properties are less well studied.

Assuming that a correct model is specified for imputation, Rubin[8] provided rules on how to do parameter estimation, construct confidence intervals, calculate P-values and carry out hypothesis testing based on Wald's method when using MI. We extend this list by developing imputation model selection criteria with the MI framework, see Section 3.

2.2 | MLE via stochastic EM

The EM algorithm[9] was designed to find MLE estimates of parameters of a parametric model in an iterative manner when the observed data are incomplete. This procedure makes use of Fisher's identity, where the maximization of the unknown observed log-likelihood is replaced with the maximization of the conditional expectation of an associated complete log-likelihood:

$$\frac{\partial l(\theta; y, r)}{\partial \theta} = \mathrm{E}_\theta \left\{ \frac{\partial l(\theta; y, z, r)}{\partial \theta} \,\middle|\, y, r \right\}.$$

An EM iteration $\theta^{(t)} \rightarrow \theta^{(t+1)}$ consists of two steps. The E-step computes the expectation of conditional complete-data log-likelihood given the observed data (with respect to the imputation model at the current estimate of parameters),

$$Q(\theta|\theta^{(t)}) = \int \log\{p(y, z, r|\theta)\}\, p(z|y, r, \theta^{(t)})\mathrm{d}z.$$

The M-step updates the estimates of parameters by maximization of the expectation function computed in the E-step,

$$\theta^{(t+1)} = \underset{\theta \in \Theta}{arg\ max}\ Q(\theta|\theta^{(t)}).$$

In situations where $Q(\theta|\theta^{(t)})$ is either analytically intractable[36] or computationally intensive,[37] it is possible to replace analytical computation of $Q(\theta|\theta^{(t)})$ by a suitable approximation to this function, commonly via simulation methods such as MCMC. In this paradigm, stochastic versions of the EM algorithm were designed to numerically compute $Q(\theta|\theta^{(t)})$ by Monte Carlo approximation. As such, the E-step in the EM algorithm simplifies to the computation of the imputation model, $p(z|y, r, \theta^{(t)})$, and simulation of the missing data $z^{(t)}$. In other words, the E-step turns into an imputation step (I-step) where

$$z^{(j)} \sim p(z|y, r, \theta^{(t)}), \quad j = 1, \ldots, M,$$

to approximate $Q(\theta|\theta^{(t)})$ as a Monte Carlo average,

$$Q(\theta|\theta^{(t)}) \simeq \frac{1}{M}\sum_{j=1}^{M} \log p(y, r, z^{(j)}|\theta).$$

A special case of the EM algorithm is the stochastic EM (StEM)[11-13] summarized in Box 1 (right). StEM, without aiming to produce any approximate computation of $Q(\theta|\theta^{(t)})$, imputes the missing data only once in the I-step until the algorithm converges to its stationary distribution, whose mean is close to the MLE.[33] The random sequence of $\{\theta^{(t)}\}$ generated by the StEM does not converge pointwise to the MLE, but, under mild conditions, does converge in distribution.[32] After the algorithm converges, multiple imputations are generated by running extra M iterations to sample from the stationary distribution, and the sample mean gives an approximate MLE of the observed likelihood.

The StEM estimator is shown to be an asymptotically normal, unbiased, and consistent estimator of θ when considering models from the exponential family.[32] Asymptotic properties of the StEM estimator are given in Wang and Robins[15] and Nielsen.[38]

2.3 | MI as stochastic EM

Studying Box 1, it is evident that the MI and StEM algorithms described above are almost identical—both iterate between imputation from the current model ("I step") and updating the imputation model ("P step"), then following convergence, both involve averaging estimates obtained from a set of M ensuing iterations. The only difference between the two algorithms is in how θ is updated at Step 2—using random draws from the current posterior, or using the maximizer of the current estimate of the likelihood function. While the former is a proper MI,[8] the latter is not since (B2, see Appendix B), and therefore (B1, see Appendix B), will no longer be satisfied.

Box 1. MI (proper) and StEM (improper) algorithms, note that they only differ at step 2

MI ("*proper*"):	StEM ("*improper*"):
0. Fix $\theta^{(0)}$ in Θ	0. Fix $\theta^{(0)}$ in Θ
1. $z^{(t+1)} \sim p(z|y, r, \theta^{(t)})$	1. $z^{(t+1)} \sim p(z|y, r, \theta^{(t)})$
2. $\theta^{(t+1)} \sim p(\theta|y, r, z^{(t+1)})$	2. $\theta^{(t+1)} = arg\ max\ p(y, r, z^{(t+1)}|\theta)$
3. Repeat steps 1 and 2 until convergence (at $t = T$)	3. Repeat steps 1 and 2 until convergence (at $t = T$)
4. Repeat steps 1 and 2 for $t = T+1, \ldots, T+M$	4. Repeat steps 1 and 2 for $t = T+1, \ldots, T+M$
5. $\hat{\theta} = \frac{1}{M}\sum_{i=T+1}^{T+M} \theta^{(i)}$	5. $\hat{\theta} = \frac{1}{M}\sum_{i=T+1}^{T+M} \theta^{(i)}$

MI views θ as random (with a prior distribution) and samples values from the observed posterior. These samples from the posterior are used to impute the missing values z, and are eventually averaged to approximate the mean of the posterior distribution of θ. StEM views θ as fixed, and at each step uses an estimate of it in the model that imputes missing values z, and eventually averages these for a final estimate of θ. From the MI point of view, this can be understood as approximating the mode of the posterior distribution with flat priors, rather than the mean. However, both algorithms

assume their estimators are asymptotically normal,[8,32] which would imply that the mean and mode would converge asymptotically.

While Rubin's combination rule[8] enables calculation of approximate standard errors when using proper MI, these are not available in the improper case. StEM nevertheless comes with recommended approaches to estimate standard errors via Louis' method, which interestingly, can be shown to have a similar form as Rubin's rule (see Appendix B).

3 | IMPUTATION MODEL SELECTION USING INFORMATION CRITERIA

The equivalence between StEM and improper MI allows standard likelihood machinery to be used to improve MI's performance. This connection changes our view of MI and provides the possibility of accessing a range of likelihood-based tools in situations where MI's performance could be improved in a likelihood-based framework.

One important gain, for example, is when a maximum likelihood framework can provide the analyst with model selection tools for choosing the best imputation model, and can enable further insights into the consequences of imputation model misspecification. In the MI literature, there are no diagnostic criteria for how to choose between imputation models. In a heuristic manner, it is recommended to either specify a multivariate normal distribution as a joint model for missing data[22] or specify a univariate conditional imputation model for each missing variable.[24] Also, it is recommended to add as many variables in the imputation model as possible, with at least as many variables as presented in the substantive analysis model of interest.[34,39] Furthermore, it is not clear how to choose between different missing data mechanisms. These are often considered untestable assumptions, which need to be determined from prior knowledge.[40] Literature suggests performing a sensitivity analysis to explore the impact of the different assumptions for missingness mechanism.[25,41] In this article, we show that available and well-known information-based criteria in the maximum likelihood literature, which enjoy good statistical properties, can be used to select an imputation model among a set of candidate imputation models.

For illustration, we will use AIC and BIC for imputation model selection, study their properties by simulation, and apply them to two real-data examples. Denote $|\theta|$ as the number of model parameters. We write $\text{AIC} = -2\sup_\theta \log p(y, r|\theta) + 2|\theta|$, and $\text{BIC} = -2\sup_\theta \log p(y, r|\theta) + \log(n) \times |\theta|$. BIC has been shown to be consistent under a broad range of settings, and we add multiple imputation models to that list in Theorem 1 below.

The equivalence of StEM and MI could in principle be used to motivate the application of any likelihood-based information criterion to imputation model selection. For example, a small-sample correction to AIC developed by Hurvich and Tsai[42] could also be applied. Cavanaugh and Shumway[43] and Ibrahim[44] developed AIC-type criteria based on the expected complete likelihood ($Q(\theta)$), which would be straightforward to approximate from imputed data. However, note that $Q(\theta)$ differs from the observed likelihood by an entropy term. In mixture models, it has been shown that if competing models differ in their entropies, complete likelihood criteria can have undesirable properties, which follow from omission of entropy from their criteria.[45] A similar problem might arise in missing data analysis, when comparing imputation models of differing complexity.

For the formulation of the next theorem, we need the notion of the most parsimonious correct model. We note that in many situations, there may be multiple correct imputation models, denoted $\mathbb{M}^{p*}$, for example, when dealing with a sequence of nested models that includes the true model.[46] In these situations, there exist multiple correct imputation models with different levels of complexity. Some of these correct imputation models may be too complex, including additional parameters with zero effect, and we define as the most parsimonious correct model $\mathbb{M}^p$ such that $|\theta_{\mathbb{M}^p}| \leq |\theta_{\mathbb{M}^{p*}}|$ for every correct $\mathbb{M}^{p*}$ model. $\mathbb{M}^p$ is typically unique but in the below we do not require it to be. In Appendix A, we give a proof of the following theorem, which shows that BIC is consistent for imputation model selection:

Theorem 1. *Suppose M_0 is the imputation model chosen by BIC and $\mathbb{M}^p$ is a finite set of the most parsimonious correct models. If Assumptions 1 to 4 (see Appendix A) are satisfied, then*

$$\Pr(M_0 \in \mathbb{M}^p) \to 1 \quad as \quad n \to \infty.$$

We can explain how the observed log-likelihood is informative about the imputation model using an approximation technique that, interestingly, is equivalent to that used in variational approximation of likelihood functions.[47] Variational approximation is a (trade-off) method for optimizing a log-likelihood function while enhancing its tractability, hence, making approximate inference for parameters of a model. In the missing data context, an incorrect imputation model

can be understood as a variational approximation to the observed log-likelihood. The approximations are indicated by the Kullback-Leibler divergence[48] of the specified imputation model from the true imputation model.

Suppose $q(z|y,r,\theta)$ is a specified imputation model and $p(z|y,r,\theta)$ is the true imputation model. Misspecification of the imputation model implies that $q(z|y,r,\theta)$ diverges from $p(z|y,r,\theta)$. The divergence of $q(z|y,r,\theta)$ from $p(z|y,r,\theta)$ is assessed by the Kullback-Leibler divergence, denoted by $KL(q||p)$, and is always nonnegative. A zero value would only occur when the imputation model is not misspecified, whereas a positive value would imply that the imputation model is misspecified, and the further $q(z|y,r,\theta)$ is from $p(z|y,r,\theta)$, the larger this divergence is, that is, the greater is $KL(q||p)$.

Following a similar approach as given in Ormerod and Wand,[47] we have

$$\begin{aligned}
\log p(y,r|\theta) &= \log p(y,r|\theta)\int q(z|y,r,\theta)\mathrm{d}z \\
&= \int q(z|y,r,\theta)\log p(y,r|\theta)\mathrm{d}z \\
&= \int q(z|y,r,\theta)\log\left(\frac{p(y,z,r|\theta)}{q(z|y,r,\theta)}\frac{q(z|y,r,\theta)}{p(z|y,r,\theta)}\right)\mathrm{d}z \\
&= \int q(z|y,r,\theta)\log\left(\frac{p(y,z,r|\theta)}{q(z|y,r,\theta)}\right)\mathrm{d}z - \int q(z|y,r,\theta)\log\left(\frac{p(z|y,r,\theta)}{q(z|y,r,\theta)}\right)\mathrm{d}z \\
&= \int q(z|y,r,\theta)\log\left(\frac{p(y,z,r|\theta)}{q(z|y,r,\theta)}\right)\mathrm{d}z + \int q(z|y,r,\theta)\log\left(\frac{q(z|y,r,\theta)}{p(z|y,r,\theta)}\right)\mathrm{d}z \\
&= \int q(z|y,r,\theta)\log\left(\frac{p(y,z,r|\theta)}{q(z|y,r,\theta)}\right)\mathrm{d}z + KL(q||p) \\
&\geq \int q(z|y,r,\theta)\log\left(\frac{p(y,z,r|\theta)}{q(z|y,r,\theta)}\right)\mathrm{d}z.
\end{aligned}$$

The maximum value of the lower bound in (2) over q is obtained when $q(z|y,r,\theta)=p(z|y,r,\theta)$ (that is, when the Kullback-Leibler divergence of q from p is at a minimum, or $KL(q||p)=0$), since $\int q(z|y,r,\theta)\log\left(\frac{p(y,z,r|\theta)}{q(z|y,r,\theta)}\right)\mathrm{d}z = \int p(z|y,r,\theta)\log\left(\frac{p(y,z,r|\theta)}{p(z|y,r,\theta)}\right)\mathrm{d}z = \log p(y,r|\theta)$. Therefore, the better the specified imputation model, the lower $KL(q||p)$, and hence, the higher $\int q(z|y,r,\theta)\log\left(\frac{p(y,z,r|\theta)}{q(z|y,r,\theta)}\right)\mathrm{d}z$. Thus, AIC and BIC based on observed log-likelihoods would be able to reflect the goodness-of-fit of the chosen imputation model.

4 | SIMULATION STUDY

We present two categories of simulation studies to investigate the performance of AIC and BIC as imputation model selection criteria where we have: (I) a univariate missing variable and (II) multivariate missing variables.

4.1 | Univariate missing variable

We consider two simulation scenarios with a univariate missing variable. In scenario (a), we fit a linear regression model with missing values in the predictor variable; and in scenario (b), we fit a log-linear (Poisson) generalized linear model (GLM) with missing values in the (count) response variable. We present scenario (b) in Appendix C2.

4.1.1 | Univariate missing predictor

We assume the linear regression model, $Y_i = 1 + X_{1i} + X_{2i} + \epsilon_i$, where Y_i, $i=1,\ldots,n$, is the response variable and the errors are $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$. Suppose that both predictors are independent and standard normal, but X_{1i} is partially observed. We imposed missingness on X_{1i} as a regression function of X_{2i}. Specifically, suppose that r_i is the missingness indicator of X_{1i} where $r_i=0$ if X_{1i} is missing and $r_i=1$ if X_{1i} is observed. Let p_i be the probability of $r_i=1$

satisfying

$$\log\left(\frac{p_i}{1-p_i}\right) = \phi_0 + \phi_1 X_{2i}.$$

Also, to investigate the consequences of overfitting an imputation model, suppose there are eight fully observed and independent auxiliary variables $X_{ki} \sim N(0,1)$ for $k = 3, \ldots, 10$.

We imposed 10%, 25% and 40% missingness in X_1 with $\phi = (2.5, 1)^\intercal$, $\phi = (1.5, 1.2)^\intercal$ and $\phi = (1, 2)^\intercal$, respectively, where ϕ is the vector $\phi = (\phi_0, \phi_1)^\intercal$. Because the missingness in X_1 only depends on X_2 the data is MAR, and thus, the missingness mechanism is ignorable.

We ran 500 simulations. We used the squared distance of parameters between the $(t+1)$th and the tth iterations as a convergence criterion for the StEM algorithm, and set the convergence threshold to 10^{-3}. Furthermore, the number of multiple imputations was set to $M = 100$. To investigate the performance of information criteria for imputation model selection, missing values $X_{mis,1}$ were imputed under two overparametrized models with one and eight extra predictors, respectively, as well as under the correct model and two wrong models (misspecified mean/distribution) as the following:

The overparametrized model (strong):

$$X_{mis,1i} \sim N(\mu_{1i}, \sigma^2)\ ,\ \mu_{1i} = \gamma_0 + \gamma_1 Y_i + \gamma_2 X_{2i} + \sum_{k=3}^{10} \gamma_k X_{ki}$$

The overparametrized model (mild):

$$X_{mis,1i} \sim N(\mu_{1i}, \sigma^2)\ ,\ \mu_{1i} = \gamma_0 + \gamma_1 Y_i + \gamma_2 X_{2i} + \gamma_3 X_{3i}$$

The correct model:

$$X_{mis,1i} \sim N(\mu_{1i}, \sigma^2)\ ,\ \mu_{1i} = \gamma_0 + \gamma_1 Y_i + \gamma_2 X_{2i}$$

The misspecified-mean model:

$$X_{mis,1i} \sim N(\mu_{1i}, \sigma^2)\ ,\ \mu_{1i} = \gamma_0 + \gamma_1 Y_i + \gamma_4 X_{3i}$$

The misspecified-distribution model:

$$\log(X_{mis,1i}) \sim N(\mu_{1i}, \sigma^2)\ ,\ \mu_{1i} = \gamma_0 + \gamma_1 Y_i + \gamma_2 X_{2i}.$$

Note that while X_1 and X_2 are independent, X_1 is actually conditionally dependent on X_2 (given Y) because Y is a function of X_1 and X_2. Therefore, the correct imputation model for X_1 is a linear function of X_2 as well as Y.

We used these five models to look at the ability of AIC and BIC to identify over-fitted imputation models as well as imputation models misspecified in two different ways: underfitted or incorrect distributional assumption.

Results in Table 1 show the proportion of times the corresponding model is chosen based on each information criterion for various sample sizes of $n = \{50, 100, 1000\}$ and missing proportions of 10%, 25%, and 40%. These results indicated that for even a small sample size of 50 and 40% missingness, both AIC and BIC are able to choose the correct model at least 86% of the time, with misspecified models selected rarely, and selected at a rate that went to zero as sample size increased. The mild overparametrized model (Overpar. mild) was chosen by AIC 2% to 5% of the time in total, irrespective of sample size. For BIC, the rate at which the correct model was chosen by BIC converged to one as the sample size increased, as expected from Theorem 1. These results align with classical results for complete data cases, where several studies have shown that AIC overfits (asymptotically)[42,49,50] while in contrast, BIC can be consistent for model selection.[51,52] A possible strategy to correct for negative bias in small sample sizes when using AIC is to use a corrected version as given in Hurvich and Tsai[42] but that would not alter the overfitting property of AIC in large samples.

TABLE 1 Proportion of times (%) the information criterion chooses the fitted imputation model for different sample sizes in 500 simulated datasets

		10%			25%			40%		
		***n*: 50**	**100**	**1000**	**50**	**100**	**1000**	**50**	**100**	**1000**
Correct	AIC	**94**	**95**	**96**	**93**	**95**	**96**	**86**	**94**	**96**
	BIC	**96**	**100**	**100**	**95**	**99**	**100**	**87**	**96**	**100**
Overpar. mild	AIC	3	5	4	2	5	4	2	4	4
	BIC	0	0	0	0	0	0	0	0	0
Overpar. strong	AIC	0	0	0	0	0	0	0	0	0
	BIC	0	0	0	0	0	0	0	0	0
Missp. mean	AIC	3	0	0	5	0	0	12	2	0
	BIC	4	0	0	5	1	0	12	4	0
Missp. dist.	AIC	0	0	0	0	0	0	0	0	0
	BIC	0	0	0	0	0	0	0	0	0

Note: The correct model (in bold) was selected most of the time in each simulation. Note that the proportion of times this model was chosen by BIC converged to one as sample size increased, as expected under Theorem 1.

The importance of choosing the correct imputation model in practice is perhaps better demonstrated by investigating its impact on post-selection inference from data. We therefore compared all methods based on the magnitude of bias and root mean square error (RMSE) for $\hat{\beta}$'s.

Figure 1 shows parallel boxplots of the relative bias of $\hat{\beta}_1$, defined here as $100 \times (\hat{\beta}_1 - \beta_1)/\beta_1$, obtained from each fitted model against increasing missing proportions and increasing sample size. Unsurprisingly, this figure showed that the misspecified models underperformed in comparison to the correct model in terms of bias as the missing proportion increased (top to bottom) and as the sample size increased (left to right). In some cases, the relative bias was as large as 40% for wrong imputation models. More interestingly, strongly overparameterized models were also downward biased (Figure 1), when missingness rate was not small (40%) and sample size was not large ($n < 100$). When sample size is small, a strongly overparameterized imputation model can lead to overfitting, and when missingness rate is large, this can have appreciable consequences in terms of bias. This result is interesting as we expected no bias but high variance from overfitting due to the commonly known bias-variance tradeoff.[53] A possible explanation for this bias comes from the measurement error modeling literature[54]—we postulate that overfitting an imputation model introduces extra variability into the imputed predictor, which can be understood as a form of measurement error. But when there is measurement error in a predictor that is not accounted for, linear model parameter estimates tend to be biased. In our simulation, this would bias estimates of β_1 toward zero, an effect known as regression attenuation.[54] This was observed in our simulations.

Table 2 shows the root mean square error (RMSE) of $\hat{\beta}_1$ of the comparing models across 500 simulations. As before, there was poor performance not only for misspecified imputation models but also for the strongly overparameterized model, when the missingness rate was not small and sample size not large (Table 2, fourth row). The mildly overparameterized model showed no detectable loss of performance.

Estimation of other regression parameters, $\hat{\beta}_0$ and $\hat{\beta}_2$ (Appendix C1), showed a similar cost of imputation model misspecification, but little impact of overfitting. We also looked at coverage probability and found that this was generally at or close to nominal levels, except in cases where the estimator was biased due to misspecification (cf. Figure 1).

We investigated how well we can avoid these issues using information criteria. We examined the impact on post-selection inference of choosing the imputation model using AIC/BIC. If making inferences about $\hat{\beta}_1$ after imputation model selection using AIC and BIC, we found negligible bias (Figure 1) and low RMSE (Table 2) in all scenarios, hence providing some protection against issues due to misspecification or overfitting of the imputation models.

Finally, while we used purpose-written StEM code to fit these models, we would expect similar results from any available multiple imputation software that was used to fit the same imputation models. To confirm this, we repeated simulations using the `Amelia`[55] and `mice`[56] `R`-packages and compared results to those seen here. `Amelia` is a bootstrapped version of the EM algorithm that assumes data are multivariate normal, and imputes the missing data by drawing from

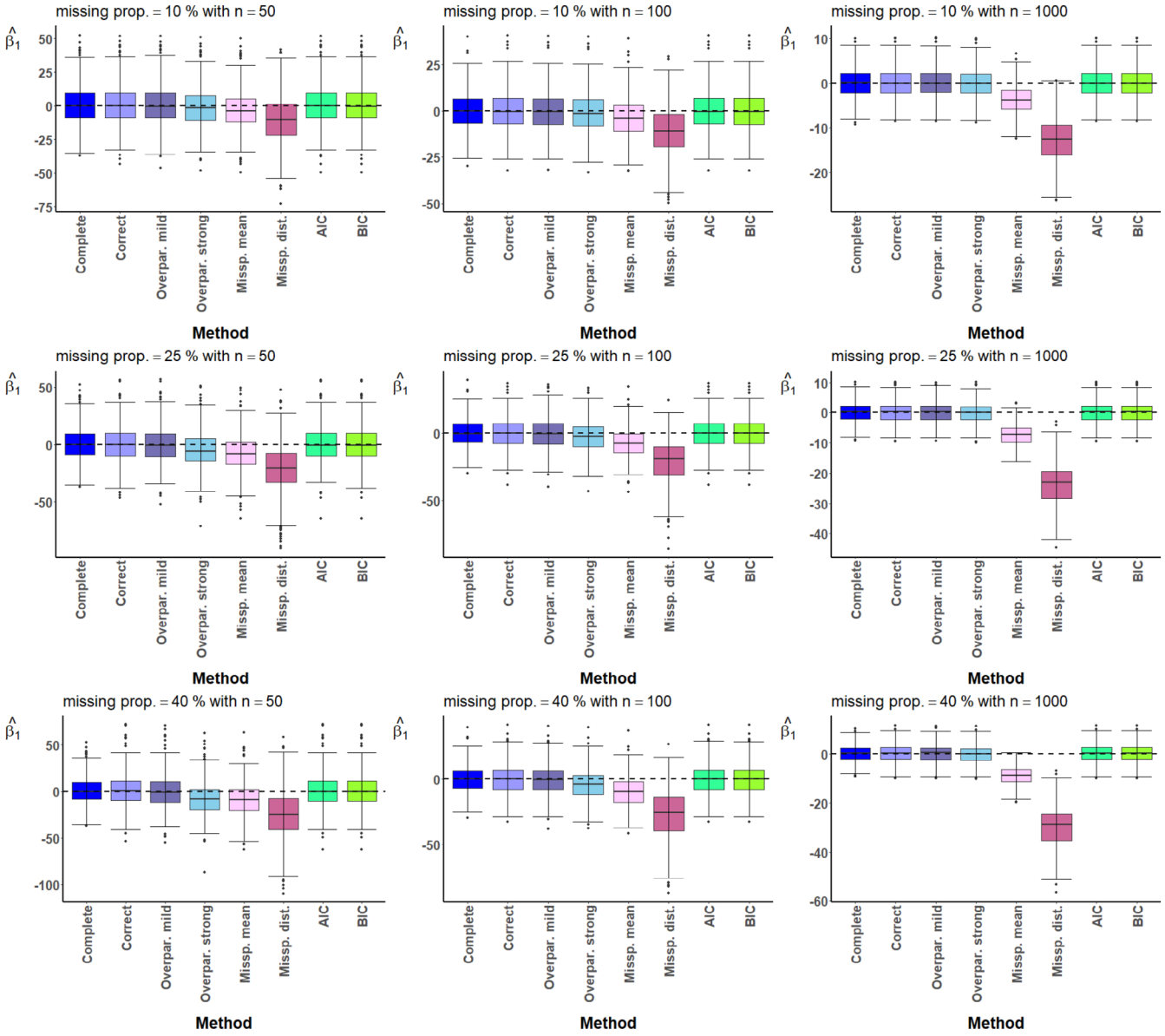

FIGURE 1 Boxplot of the relative bias of $\hat{\beta}_1$ for different methods. Each method is compared with the original dataset (Complete) based on the accuracy of their estimators as the missing proportion and the sample size increase from the top-left corner to the bottom-right. Note that AIC/BIC refer to the model chosen by AIC/BIC, respectively, among the comparing imputation models [Colour figure can be viewed at wileyonlinelibrary.com]

their posterior distribution given the estimated parameters. The `mice` package with the `mice.impute.norm.nob()` function was used to impute X_1 using improper imputation by linear regression (MICE). We saw similar patterns for `Amelia` and `MICE` to those presented here (see Appendix C1).

4.2 | Multivariate missing variable

Next, we investigated the case where missingness was of multiple dimension, and the response was non-normal (here we considered binary data). The purpose of this simulation study is to investigate the performance of information criteria for choosing between MAR and MNAR models with MI. Following example 4.1.2 from Ibrahim,[57] we now suppose that the response variables Y_i, $i = 1, \ldots, n$, are independent fully observed Bernoulli random variables and we are interested in the analysis of a logistic regression (GLM) model $\mathrm{E}(Y_i|X_{1i}, X_{2i}, \beta) = \exp(X_i\beta)/\{1 + \exp(X_i\beta)\}$ where $\beta = (\beta_0, \beta_1, \beta_2)^\intercal$ and

TABLE 2 RMSE (×1000) of $\hat{\beta}_1$ for different imputation methods across 500 simulations, compared with the original dataset (Complete)

	10%			25%			40%		
	n=50	n=100	n=1000	n=50	n=100	n=1000	n=50	n=100	n=1000
Complete	146	103	32	146	103	32	146	103	32
Correct	152	107	32	159	111	34	171	115	36
Overpar. mild	152	106	32	159	111	34	171	115	36
Overpar. strong	152	108	33	168	114	34	195	122	36
Missp. mean	152	114	50	174	135	81	195	153	98
Missp. dist.	202	172	136	305	268	248	373	339	310
AIC	152	107	32	161	111	34	174	115	36
BIC	152	107	32	161	111	34	174	115	36

Note: The true value of the slope coefficient is 1.

$X_i = [1\ X_{1i}\ X_{2i}]$. Suppose that the predictors, $X_i = (X_{1i}, X_{2i})$, $i = 1, \ldots, n$, are partially observed variables from a bivariate normal distribution $N_2(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 0.25 & 0.125 \\ 0.125 & 0.25 \end{pmatrix}.$$

Also, suppose that $r_i = (r_{i1}, r_{2i})$ is the missingness indicator vector of (X_{1i}, X_{2i})—for example, $r_i = (0,0)$ if X_{1i} and X_{2i} are both missing and $r_i = (0,1)$ if X_{1i} is missing and X_{2i} is observed. Let p_1 and p_2 be the probabilities of $r_{1i} = 1$ and $r_{2i} = 1$, respectively, which satisfy $\text{logit}(p_1) \equiv \log(p_1/(1-p_1)) = \phi_{10} + \phi_{11}X_{1i} + \phi_{12}X_{2i} + \phi_{13}Y_i$ and $\text{logit}(p_2) = \phi_{20} + \phi_{21}X_{1i} + \phi_{22}X_{2i} + \phi_{23}Y_i + \phi_{24}r_{1i}$.

We are interested in the comparison of the following two imputation models:

Model 1: $p(X_{mis,i} \mid X_{obs,i}, Y_i, r_i)$ and
Model 2: $p(X_{mis,i} \mid X_{obs,i}, Y_i)$,

where Model 1 assumes that the data are MNAR and missingness is nonignorable whereas Model 2 assumes that data are MAR and ignores the missingness.

We ran 500 simulations with $n = 250$, $\beta = (1, 1, -1)^\intercal$, $\phi_1 = (1, -1, 1, 1)^\intercal$ and $\phi_2 = (1, -1, 1, 1, -0.5)^\intercal$ where on average we obtained about 19% missing data in X_1, 25% in X_2 and 6% in both. Also, to draw imputations at the ith iteration of StEM, we used Metropolis Hastings within a Gibbs sampler[58] to generate a sample from

$$p(X_{mis,i}|X_{obs,i}, Y_i, r_i, \gamma_1^{(t)}) \propto p(r_i|Y_i, X_{mis,i}, X_{obs,i}, \phi^{(t)})p(Y_i|X_{mis,i}, X_{obs,i}, \beta^{(t)})p(X_{mis,i}, X_{obs,i}|\alpha^{(t)})$$

for Model 1, and a sample from

$$p(X_{mis,i}|X_{obs,i}, Y_i, \gamma_2^{(t)}) \propto p(Y_i|X_{mis,i}, X_{obs,i}, \beta^{(t)})p(X_{mis,i}, X_{obs,i}|\alpha^{(t)})$$

for Model 2 where $\alpha = (\mu, \Sigma)$, $\gamma_1 = (\phi, \beta, \alpha)$ and $\gamma_2 = (\beta, \alpha)$.

Once again, we used the squared distance of the parameters between the $(t+1)$th and tth iterations as the convergence criterion and set the convergence threshold to 10^{-4} with the number of multiple imputations set to $M = 100$.

Table 3 shows the proportion of times each model was chosen based on the corresponding information criterion. This table shows that AIC and BIC were able to choose the correct model (Model 1) 99.4% and 95.2% of the times, respectively. These results are consistent with the previous simulation results (cf. Section 4.1.1) as well as with our theoretical result (cf. Section 3), indicating that in situations where AIC and BIC are applicable, they perform satisfactory for imputation model selection.

TABLE 3 Proportion of times (%) the information criterion chooses the corresponding model across 500 simulated datasets

	Model 1	Model 2
AIC	**99.4**	0.6
BIC	**95.2**	4.8

Note: Data were generated under Model 1, which assumes MNAR, whereas Model 2 assumed data were MAR. Note that both approaches were able to recover the correct model with high probability.

TABLE 4 $\overline{\beta}_j$ and mean square error (in parenthesis) for different imputation models across 500 simulations

	$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -1$
Model 1	1.008 (0.027)	1.014 (0.083)	−1.018 (0.084)
Model 2	1.521 (0.342)	1.288 (0.177)	−1.295 (0.186)

One again, we investigate the impact of imputation model selection, here between MNAR and MAR models, on post-selection inference from data. Table 4 shows the sample average of estimates of β_j, denoted as $\overline{\beta}_j$, $j = 0, 1, 2$, for the corresponding imputation model averaged over 500 simulations *and* its mean square error, $\text{MSE}_j = (\overline{\beta}_j - \beta_j)^2 + s_j^2$, where s_j is the simulated standard error of $\overline{\beta}_j$. This table shows that parameter estimates can be heavily biased if the wrong imputation model is used (Model 2), with much larger mean squared errors.

5 | EXAMPLES

We give two real-data examples in health research studies where missing values arise. The aim in both examples is to select the most appropriate imputation model, using the tools presented in Section 3.

5.1 | Survival of infants

The following example is taken from example 9.8 of Little and Rubin[27] where a 2^3 contingency table on the survival of infants is analysed in the presence of missing data. Suppose we have three dichotomous variables of Prenatal care ($i = \{Less, More\}$), Survival ($j = \{Died, Survived\}$), and Clinic ($k = \{A, B\}$).

Let n_{ijk} denote the total count of the ijkth cell, and $n = \sum_i \sum_j \sum_k n_{ijk}$ be the total sample size. Table 5 is (artificially) partially classified for about 26% of the data (255 cases out of 970) where we observed only $n_{ij.}$ instead of n_{ijk}. The remainder of the data (715 cases out of 970) are completely observed. Also, let π_{ijk} denote classification probability for the ijkth cell. The response variable is the cell counts, n_{ijk}. Subject to missingness, these cell counts can be modeled by assuming different associations between the three predictors: Survival (S), Prenatal care (P), and Clinic (C).

Assuming that these data are MAR and the missingness is ignorable, Little and Rubin[27] applied the EM algorithm to fit various models, such as $\{SC, SP, PC\}$, $\{SP, SC\}$, and $\{SC, PC\}$ where, for example, $\{SC\}$ denotes a model with all the main effects of Survival, Prenatal, and Clinic *and* the interaction effect between Survival and Clinic. The goodness-of-fit using likelihood-ratio tests was then assessed for each candidate model. Meng and Rubin[59] applied Bayesian MI with a full (saturated) model where they tested the null models $\{SC, PC\}$ and $\{S, P, C\}$ (a main effects only model) against the full model using their proposed pooled likelihood-ratio test developed for MI. Both approaches concluded that Survival is related to Clinic, but conditional on Clinic, Survival, and Prenatal care are independent indicating that $\{SC, PC\}$ is the best parsimonious fitted model.

This example can also be viewed as a problem of imputation model selection, where the response variable is the missing variable. Thus, the model that we choose to fit to the data can be used as the imputation model in the MI algorithm. As such, we will have imputation model candidates $\{SC, SP, PC\}$, $\{SP, SC\}$, $\{SC, PC\}$, and so on.

We fitted five competing imputation models as follows:

TABLE 5 Survival of infants taken directly from example 9.8 of Little and Rubin[27]

		Survival (S)		
Clinic (C)	**Prenatal care (P)**	***Died***	***Survived***	
A	*Less*	3	176	
	More	4	293	
B	*Less*	17	197	
	More	2	23	complete = 715
?	*Less*	10	150	
	More	5	90	partial = 255

TABLE 6 Information criteria for each candidate imputation model for the survival of infants example

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
AIC	27.02	**25.12**	33.29	156.71	168.28
BIC	63.60	**57.13**	65.30	188.71	191.14

Note: Both AIC and BIC favored Model 2, $\{SC, PC\}$, in line with previous analyses.[27,59]

Model 1: $\{SC, SP, PC\}$,
Model 2: $\{SC, PC\}$,
Model 3: $\{SP, PC\}$,
Model 4: $\{SP, SC\}$,
Model 5: $\{S, P, C\}$.

To demonstrate the model fitting procedure, the StEM/improper MI algorithm can be applied to this problem by the following iterative steps:

1. Estimate initial value $\hat{\pi}_{ijk}^{(0)}$ based on the observed counts.
2. For tth iteration, $t = 1, \ldots, T$,

- Simulate $\hat{n}_{ijk}^{(t)}$ from $Bin(n_{ij.}, \hat{\pi}_{ijk}^{(t-1)})$.
- Re-estimate π_{ijk} as $\hat{\pi}_{ijk}^{(t)} = arg\ max\ \ell(\pi)$, where $\ell(\pi) = \sum_i \sum_j \sum_k n_{ijk} \log \pi_{ijk}$ is the complete log-likelihood of a multinomial model. For instance, under Model 2, imputations for the partially classified counts of the $ij1$th cell are drawn from $\hat{n}_{ij1} \sim Bin(n_{ij.}, \hat{\pi}_{ij1})$ where $\hat{\pi}_{ij1} = \hat{\mu}_{ij1}/(\hat{\mu}_{ij1} + \hat{\mu}_{ij2})$ and $\hat{n}_{ij2} = n_{ij.} - \hat{n}_{ij1}$.

3. Calculate the AIC/BIC for the fitted model using criteria presented in Section 3.

Table 6 shows the AIC and BIC for the above-mentioned imputation models based on $M = 100$ multiple imputations. Also, the convergence threshold for the algorithm was set to 10^{-4}. AIC and BIC both favored Model 2, in line with the results of Little and Rubin[27] and Meng and Rubin.[59]

Table 7 shows the estimated cell probabilities for the above-mentioned imputation models. These estimates differ under each imputation model. For example, the percentage of infants dying with less prenatal care in clinic A, $\pi_{111} \times 100$, is estimated at 0.45 and 0.49 under the correct imputation model and under the overfitted imputation model, respectively. However, this percentage is estimated to be much higher under Model 3 (at $\hat{\pi}_{111} \times 100 = 0.82$), and even higher under Model 4 (at $\hat{\pi}_{111} \times 100 = 1.41$) and under Model 5 (at $\hat{\pi}_{111} \times 100 = 1.60$). This result shows that the cell probabilities are estimated more similarly by the correct model $\{SC, PC\}$ and under the overfitted model $\{SC, SP, PC\}$. However, there is clear a difference between $\hat{\pi}_{ijk}$s under Model 3, Model 4, and Model 5 and $\hat{\pi}_{ijk}$ under the correct imputation model $\{SC, PC\}$. The estimated cell probabilities in Table 7 can be compared with their ML estimates via the EM algorithm obtained in Little and Rubin,[27] table 9.9.

TABLE 7 Estimated cell probabilities $\hat{\pi}_{ijk} \times 100$ for candidate imputation models in Survival of infants data

			Survival (S)	
	Clinic (C)	**Prenatal care (P)**	**Died**	**Survived**
Model 1: $\{SC, SP, PC\}$	A	Less	0.45	25.35
		More	0.79	38.78
	B	Less	2.64	28.57
		More	0.34	3.07
Model 2: $\{SC, PC\}$	A	Less	0.49	25.27
		More	0.76	38.79
	B	Less	2.68	28.57
		More	0.30	3.15
Model 3: $\{SP, PC\}$	A	Less	0.82	36.66
		More	0.30	28.46
	B	Less	2.27	17.26
		More	0.83	13.40
Model 4: $\{SP, SC\}$	A	Less	1.41	24.64
		More	1.05	38.59
	B	Less	1.68	29.27
		More	0.09	3.23
Model 5: $\{S, P, C\}$	A	Less	1.60	36.34
		More	1.21	27.41
	B	Less	0.81	18.26
		More	0.09	3.27

Note: These estimates differ under each model. Although cell probabilities are estimated more similarly under the overfitted model $\{SC, SP, PC\}$ and the correct model $\{SC, PC\}$, there is a clear difference between $\hat{\pi}_{ijk}$s under Models 3, 4, and 5 and $\hat{\pi}_{ijk}$ under the correct imputation model $\{SC, PC\}$.

5.2 | Pima Indian women

In our next example, we analyzed data collected on 768 Pima Indian women of at least 21 years of age living near Phoenix, Arizona who were tested for diabetes according to World Health Organization criteria. These data were collected by the U.S. National Institute of Diabetes and Digestive and Kidney Diseases, and are available in the `mlbench` R-package. These data were analyzed in Miller et al[60] and further in Smith et al[61] and Ripley.[62] The binary response variable indicates whether or not diabetes was diagnosed within 5 years of the examination (positive/negative). The explanatory variables are

1. number of pregnancies,
2. plasma glucose concentration at 2 hours in an oral glucose tolerance test,
3. diastolic blood pressure (mm Hg),
4. triceps skin fold thickness (mm),
5. 2-hour serum insulin (μU/ml),
6. body mass index (kg/m^2),
7. diabetes pedigree function,
8. age (years).

There are five values missing (0.6%) for plasma glucose concentration, 11 values (1.4%) for body mass index, 35 values (4.6%) for diastolic blood pressure, 227 values (30%) for triceps, and 374 values (49%) for serum insulin. We fitted a logistic regression generalized linear model assuming three different imputation models. These are

TABLE 8 Information criterion for different imputation models fitted to the Pima Indian women data

	Model 1	Model 2	Model 3
AIC	**7231.142**	7660.071	10544.15
BIC	**7714.096**	8026.931	10818.13

Note: The data strongly favored Model 1, suggesting a non-ignorable missing data mechanism.

TABLE 9 Estimates of regression parameters and their standard errors (in parentheses) for different imputation models in the Pima Indian women example

	Intercept	Pregnancy	Glucose	Pressure	Triceps	Insulin	BMI	Pedigree	Age
Model 1	−0.912 (0.102)	0.282 (0.113)	0.918 (0.143)	−0.034 (0.107)	0.163 (0.152)	0.321 (0.208)	0.503 (0.139)	0.328 (0.098)	0.284 (0.120)
Model 2	−0.890 (0.100)	0.290 (0.113)	1.080 (0.185)	−0.124 (0.110)	0.088 (0.199)	0.022 (0.267)	0.606 (0.179)	0.322 (0.098)	0.315 (0.119)
Model 3	−1.055 (0.154)	0.076 (0.187)	1.055 (0.178)	−0.025 (0.146)	0.114 (0.180)	0.084 (0.181)	0.449 (0.190)	0.407 (0.147)	0.628 (0.211)

Note: Pregnancy, insulin, pedigree, and age are log-transformed.

Model 1: the missingness mechanism is nonignorable (MNAR),
Model 2: the missingness mechanism is ignorable (MAR),
Model 3: complete case estimation (MCAR).

Using the StEM/improper MI via Metropolis Hastings within a Gibbs sampler[58] similar to Section 4.2, we assume that under Model 1, the missingness mechanism is $p(r_i \mid X_{mis,i}, X_{obs,i}, Y_i)$, and under Model 2, the missingness mechanism is $p(r_i \mid X_{obs,i}, Y_i)$ where r_i is the joint missingness indicator and X_i is the joint representation of all explanatory variables. Under the complete case model, we assume that r_i is independent of X_i and Y_i.

Table 8 shows the AIC and BIC values compared for each fitted model where both information criteria choose Model 1 over the other models. This strongly suggests that the missingness mechanism is nonignorable. Also, Table 9 shows the estimates of regression parameters for each imputation model. Note that results given in this table are based on log-transformations of insulin, pregnancy, pedigree and age as well as standardization on all predictors, and its interpretation here is only used for drawing comparison between the three imputation models. We see that quite different results are obtained for these three models. For example, the regression parameter for 2-hour serum insulin is estimated at 0.321 for Model 1, 0.022 for Model 2, and 0.084 for Model 3. This result suggests how different conclusions can be drawn depending on whether we include the missingness mechanism in the model and whether or not it is worth doing so.

Note that the nonignorability of the missingness mechanism necessarily relies on parametric assumptions. In this example, we have assumed that $[r_i|X_i, Y_i]$ follows a multivariate Bernoulli distribution with a logit link function. Also, for Model 1, we assume nonignorability of missingness mechanism for all the missing variables. A further investigation of different parametric assumptions for r_i (or whether nonignorability assumption can be relaxed for some of the missing variables) may be carried out if one has a reasonable argument for considering these in the study. For example, a visual inspection of the density plots of the observed and imputed data of triceps skin fold thickness and 2-hour serum insulin for Model 1 in Figure 2 might suggest further investigations into whether missingness mechanism for triceps skin fold thickness can be ignored.

6 | DISCUSSION

By exploiting the connection between MLE and MI, our findings show that we can diagnose imputation models for misspecification using standard likelihood-based information criteria such as AIC and BIC. Moreover, for the first time, we have provided insights into imputation model misspecification via variational approximation. Furthermore, we examined some theoretical properties of BIC. We showed that BIC is consistent for the correct imputation model which also holds when the correct model is missing not at random. We analyzed two health research datasets to demonstrate the method's flexibility for imputation model selection in the presence of missing data and the use of multiple imputation. For these examples, we evaluated the post-imputation effect of imputation model selection on parameter estimates. We conducted three simulation studies and showed that imputation model misspecification and overfitting can have substantial costs in terms of bias and efficiency of parameter estimation, a problem that can often be avoided using model selection. These

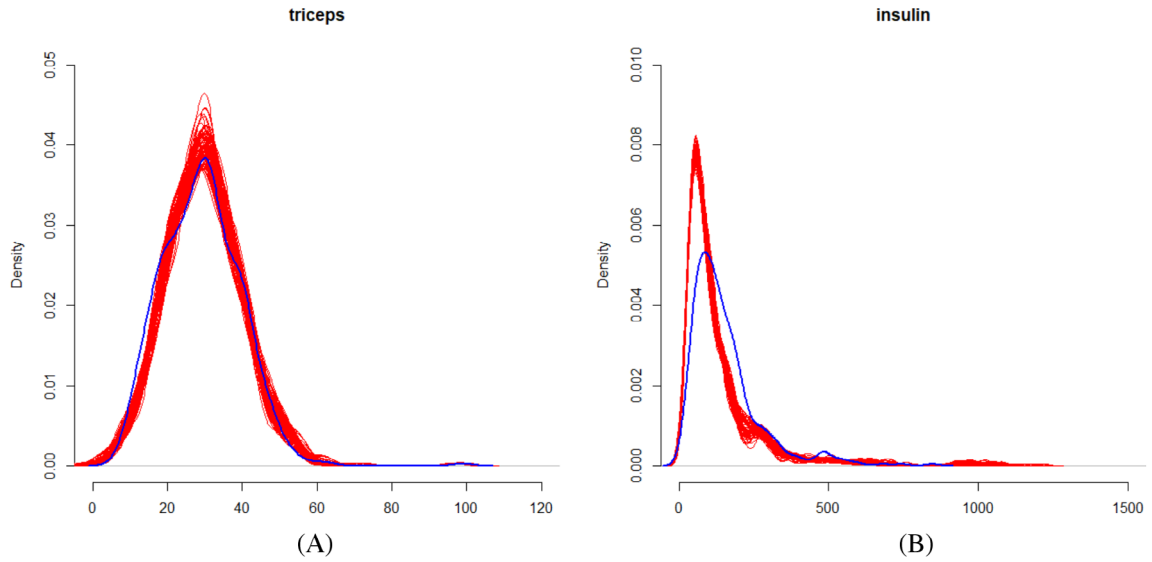

FIGURE 2 Density curve plots of, A, observed triceps skin fold thickness (blue) and multiple imputed values of triceps skin fold thickness (red) and, B, observed 2-hour serum insulin (blue) and multiple imputed values of 2-hour serum insulin (red) for Model 1 in Pima Indian Women example. Similarity between the observed curve and imputed curves in (A) suggests that triceps skin fold thickness might be missing at random. Difference between the observed curve and imputed curves in (B) suggests that 2-hour serum insulin is missing not at random [Colour figure can be viewed at wileyonlinelibrary.com]

results are not surprising—the costs of misspecification and overfitting models are well-known—but it is perhaps surprising that this has been little considered previously in the context of choosing an *imputation model*, where the ideas apply equally well, as does the notion of using model selection to guard against such problems.

A simulation result that we did not anticipate is that overfitting the imputation model had costs not just in loss of efficiency, but also in bias. This is an unusual finding because overfitting a model is conventionally thought to not bias predictions, rather to increase their variance (the "bias-variance tradeoff"[53]). We speculate that bias arose in our simulations via a similar mechanism to regression attenuation in measurement error modeling.[54] Specifically, overfitting the imputation model would lead to larger variance in the imputed missing values of predictors, which could be interpreted as a form of measurement error on predictors. Unaccounted-for measurement error in predictors tends to bias coefficient estimators,[54] and in our simulation this would bias the estimator of β_1 toward zero, as we observed.

While our simulations suggested negligible loss of efficiency when mildly overparametrizing the imputation model (with just one unneeded predictor included), this need not be true in general. It became clear in our simulation work that the sensitivity to choice of imputation model varied considerably with the context, and some situations are much more sensitive to imputation method than others.

The fitted models used in this article were quite simple so it would be of interest to see how the criteria perform for more complicated scenarios such as in high-dimensional settings where the number of variables or components available for use in the analysis is larger than the number of observations, or when including random effects in the model. That said, our approach is very general and applies to any scenario where MI is used on a regression model that is fitted by ML for complete data, and thus, we do not anticipate any difficulties extending simulation studies to other settings such as mixed models, beyond the consideration of which information criterion to use.[63]

The application of our findings may be extended to other areas where MI's performance could be improved in a likelihood based framework. For example, suppose that we would like to predict myocardial infarction in patients with observed blood pressure, body mass index, age, and gender, but cholesterol level is subject to missingness. MI's approach based on Rubin's rules does not provide any guidance on how to approach prediction in the presence of missing data, and most of the methods suggested in the MI literature are ad hoc.[64] However, there is a clear guidance on how to carry out prediction in a likelihood based framework and the close connection between MI and StEM could provide clarity in this field.

Another example of where likelihood based inference might improve MI's performance is in hypothesis testing. In the MI literature, hypothesis testing based on multiple imputed datasets were proposed to obtain a modified Wald test statistic.[8] Subsequent work by Meng and Rubin[59] developed a pooling procedure for a likelihood ratio test with MI for nested models, using the asymptotic relationship between the Wald test and likelihood ratio test statistics. Although this approach works well, it can be quite cumbersome to implement in practice. Using connections with maximum

likelihood, a likelihood ratio statistic could be directly constructed for MI. Our preliminary work suggests this has improved performance over other statistics in the MI literature.

We primarily focused on the StEM algorithm in this article, however, an alternative Monte-Carlo approximation to the EM algorithm, known as MCEM,[10] could similarly be used. This algorithm obtains an (approximate) MLE by averaging likelihood estimates across the multiple imputations, then maximizing, as opposed to StEM which maximizes on each imputation and then averages estimates. The MCEM algorithm is more efficient than the StEM algorithm for finite sample sizes and for finite number of imputations.[38] This is due to the fact that StEM loses some efficiency due to the maximize-then-average strategy. Therefore, in situations where there exists little concern for computational efficiency, it would be beneficial to apply MCEM instead of StEM without having to compromise any of the results discussed in this article.

Finally, in this article, we focused on using a fixed value for θ rather than random θ. As such, there is the question of whether there are any implications from using improper rather than proper MI, where we optimize for θ rather than sampling it from a distribution (see Box 1, Step 2). While this was done in order to make a connection to maximum likelihood, there are few issues anticipated in extending our results to the random θ scenario of proper MI. The reason for this is that the inference machinery behind likelihood-based inference uses asymptotic theory as $n \to \infty$, and in this setting, the posterior distribution for θ will typically converge in probability to its optimized value, hence we anticipate negligible effect of sampling θ from its posterior rather than plugging in its optimized value. One possible adjustment to the proposed approaches when θ is random is to instead use Bayesian information criteria, such as the deviance information criterion.[65] Future work could study whether similar arguments apply for random parameters.

ACKNOWLEDGEMENT

This work was supported by an Australian Government Research Training Program Scholarship.

REFERENCES

1. Carpenter J, Pocock S, Johan LC. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. *Stat Med*. 2002;21(8):1043-1066.
2. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: National Academies Press; 2010.
3. Karon JM, Song R, Brookmeyer R, Kaplan EH, Hall HI. Estimating HIV incidence in the United States from HIV/AIDS surveillance data and biomarker HIV test results. *Stat Med*. 2008;27(23):4617-4633.
4. Stirrup OT, Babiker AG, Carpenter JR, Copas AJ. Fractional Brownian motion and multivariate-t models for longitudinal biomedical data, with application to CD4 counts in HIV-positive patients. *Stat Med*. 2016;35(9):1514-1532.
5. Chen HY, Little RJ. Proportional hazards regression with missing covariates. *J Am Stat Assoc*. 1999;94(447):896-908.
6. White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med*. 2009;28(15):1982-1998.
7. Laird NM. Missing data in longitudinal studies. *Stat Med*. 1988;7(1-2):305-315.
8. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley; 1987.
9. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc Ser B Methodol*. 1977;39(1):1-38.
10. Wei GC, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J Am Stat Assoc*. 1990;85(411):699-704.
11. Broniatowski M, Celeux G, Diebolt J. Reconnaissance de melanges de densites par un algorithme d'apprentissage probabiliste. *Data Anal Inform*. 1983;3:359-374.
12. Celeux G, Diebolt J. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput Stat Quart*. 1985;2(1):73-82.
13. Celeux G, Diebolt J. A probabilistic teacher algorithm for iterative maximum likelihood estimation. *Classification and Related Methods of Data Analysis*. Amsterdam, Netherlands: North Holland; 1987:617-623.
14. Honaker J, King G. What to do about missing values in time-series cross-section data. *Am J Pol Sci*. 2010;54(2):561-581.
15. Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. *Biometrika*. 1998;85(4):935-948.
16. Biscarat JC, Celeux G, Diebolt J. *Stochastic Versions of the EM Algorithm. Technical Report*. Seattle: Washington University Department of Statistics; 1992.
17. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Stat Med*. 2008;27(17):3227-3246.
18. May M, Boulle A, Phiri S, et al. Prognosis of patients with HIV-1 infection starting antiretroviral therapy in sub-Saharan Africa: a collaborative analysis of scale-up programmes. *The Lancet*. 2010;376(9739):449-457.
19. Schomaker M, Heumann C. Model selection and model averaging after multiple imputation. *Comput Stat Data Anal*. 2014;71:758-770.
20. Fay RE. A design-based perspective on missing data variance. Paper presented at: Proceedings of the 1991 Annual Research Conference; 1991:429-440; US Bureau of the Census, Washington, DC.

21. Fay RE. Alternative paradigms for the analysis of imputed survey data. *J Am Stat Assoc*. 1996;91(434):490-498.
22. Schafer JL. *Analysis of Incomplete Multivariate Data*. London, UK: Chapman & Hall/CRC Press; 1997.
23. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol*. 2009;60:549-576.
24. van Buuren S. *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC Press; 2012.
25. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat Methods Med Res*. 2007;16(3):259-275.
26. Little RJA. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc*. 1993;88(421):125-134.
27. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York, NY: John Wiley & Sons; 2002.
28. Andridge R, Thompson KJ. Using the fraction of missing information to identify auxiliary variables for imputation procedures via proxy pattern-mixture models. *Int Stat Rev*. 2015;83(3):472-492.
29. Seaman S, Galati J, Jackson D, Carlin J. What is meant by "missing at random"? *Stat Sci*. 2013;28(2):257-268.
30. Mealli F, Rubin DB. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*. 2015;102(4):995-1000.
31. Doretti M, Geneletti S, Stanghellini E. Missing data: a unified taxonomy guided by conditional independence. *Int Stat Rev*. 2018;86(2):189-204.
32. Diebolt J, Celeux G. Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Stoch Models*. 1993;9(4):599-613.
33. Diebolt J, Ip EH. Stochastic EM: method and application. *Markov Chain Monte Carlo in Practice*. New York, NY: Springer; 1996:259-273.
34. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc*. 1996;91(434):473-489.
35. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.
36. McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *J Am Stat Assoc*. 1998;92(437):162-170.
37. Ng SK, McLachlan GJ. On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Stat Comput*. 2003;13(1):45-55.
38. Nielsen SF. The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*. 2000;6(3):457-489.
39. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330-351.
40. Mohan K, Pearl J. Graphical models for processing missing data. *J Am Stat Assoc*. 2020.
41. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
42. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. *Biometrika*. 1989;76(2):297-307.
43. Cavanaugh JE, Shumway RH. An Akaike information criterion for model selection in the presence of incomplete data. *J Stat Plan Infer*. 1998;67(1):45-65.
44. Ibrahim JG, Zhu H, Tang N. Model selection criteria for missing-data problems using the EM algorithm. *J Am Stat Assoc*. 2008;103(484):1648-1658.
45. Hui FK, Warton DI, Foster SD. Order selection in finite mixture models: complete or observed likelihood information criteria? *Biometrika*. 2015;102(3):724-730.
46. Wang L, Qu A. Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J Royal Stat Soc Ser B Methodol*. 2009;71(1):177-190.
47. Ormerod JT, Wand MP. Explaining variational approximations. *Am Stat*. 2010;64(2):140-153.
48. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat*. 1951;22(1):79-86.
49. Shibata R. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*. 1976;63(1):117-126.
50. Bozdogan H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*. 1987;52(3):345-370.
51. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.
52. Nishii R. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann Stat*. 1984;12(2):758-765.
53. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Comput*. 1992 Jan;4(1):1-58.
54. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, FL: CRC Press; 2006.
55. Honaker J, King G, Blackwell M, Amelia II. A program for missing data. *J Stat Softw*. 2011;45(7):1-47.
56. Buuren SV, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2010;45:1-68.
57. Ibrahim JG, Lipsitz SR, Chen MH. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J Royal Stat Soc Ser B Methodol*. 1999;61(1):173-190.
58. Tierney L. Exploring posterior distributions using Markov chains. In: Keramidas E, ed. *Proceedings of the 23rd Symposium Interface Computer Science and Statistics*. Fairfax Station: Interface Foundation; 1991:563-570.
59. Meng XL, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*. 1992;79(1):103-111.
60. Miller M, Burch TA, Bennett PH, Steinber AG. Prevalence of diabetes mellitus in American Indians – results of glucose tolerance tests in Pima Indians of Arizona. *Diabetes*. 1965;14(7):439-440.
61. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Paper presented at: Proceedings of the Annual Symposium on Computer Applications in Medical Care, Washington, DC: IEEE Computer Society Press; 1988:261–265.
62. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge, MA: Cambridge University Press; 1996.

63. Vaida F, Blanchard S. Conditional Akaike information for mixed-effects models. *Biometrika*. 2005;92(2):351-370.
64. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biom J*. 2015;57(4):614-632.
65. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *J Royal Stat Soc Ser B Methodol*. 2002;64(4):583-639.
66. Sen PK, Singer JM. *Large Sample Methods in Statistics: An Introduction with Applications*. Vol 25. Boca Raton, FL: CRC Press; 1994.
67. Louis TA. Finding the observed information matrix when using the EM algorithm. *J Royal Stat Soc Ser B Methodol*. 1982;44(2):226-233.
68. von Hippel PT. The bias and efficiency of incomplete-data estimators in small univariate normal samples. *Sociol Methods Res*. 2012;42(4):531-558.

How to cite this article: Noghrehchi F, Stoklosa J, Penev S, Warton DI. Selecting the model for multiple imputation of missing data: Just use an IC!. *Statistics in Medicine*. 2021;40:2467–2497. https://doi.org/10.1002/sim.8915

APPENDIX A. PROOF OF THEOREM 1

Let $y=(y_1,y_2,\ldots,y_{n-m})$ and $z=(z_0,\ldots,z_m)$, where $n\in\mathbb{N}$ and $m\in\{0,1,\ldots,n-1\}$, and denote $r=(r_1,\ldots,r_n)$ as the missingness indicator. Suppose that $y_1,y_2,\ldots,y_{n-m}$ and $r_1,r_2,\ldots,r_n$ are independent observations. Let $\mathbb{S}$ denote the set of all possible imputation models, $\mathbb{S}=\{M_k=q(z|y,r,\theta_k);k=0,1,2,\ldots,K\}$ with $|\theta_k|$ denoting the total number of parameters for model M_k, and K being the number of competing imputation models. Also, define a subset $\mathbb{M}\subset\mathbb{S}$ of models which minimize $KL(q||p)$. Furthermore, $\mathbb{M}^p\subset\mathbb{M}$ denotes the finite subset of the most parsimonious correct imputation models, $\mathbb{M}^p=\{M_k\in\mathbb{M}:|\theta_k|\leq|\theta_{k^*}|,\forall M_{k^*}\in\mathbb{M}\}$. For simplicity, we denote $q_k(z)=q(z|y,r,\theta_k)$ throughout this section.

To prove consistency of BIC for imputation model selection, we require a set of regularity conditions that consist of the following assumptions:

Assumption 1. Observations $z_0,z_1,\ldots,z_m$ are independent, and densities $q_k(z_i)$ exist.

Assumption 2. $m=o_p(n)$.

Assumption 3. The derivatives of the likelihood function $\int q_k(z)\log\left(p(y,z,r|\theta_k)/q_k(z)\right)\mathrm{d}z$ up to order three exist w.r.t. θ_k, and are continuous and uniformly bounded for all $\theta_k\in\Theta^{(k)}$.

Assumption 4. The derivatives of observed likelihood function up to order three exist w.r.t. θ and are continuous and uniformly bounded for all $\theta\in\Theta$.

We also need Lemmas 1 and 2:

Lemma 1. *Let $q_t(z)\in\mathbb{M}$ be an arbitrary correct imputation model and $q_w(z)\in\mathbb{M}^c$ be an arbitrary wrong imputation model where $\mathbb{M}^c$ denotes the complement of $\mathbb{M}$. Then, we have*

$$\int q_t(z)\log\left(\frac{p(y,z,r|\theta_t)}{q_t(z)}\right)\mathrm{d}z>\int q_w(z)\log\left(\frac{p(y,z,r|\theta_w)}{q_w(z)}\right)\mathrm{d}z.$$

Proof. The proof is straightforward by using the relationship between the imputation model and observed log-likelihood in (2). Suppose $q_k(z)$ is any specified imputation model in $\mathbb{S}$. Let Q be a class of lower bounds based on various imputation models,

$$Q=\left\{\int q_k(z)\log\left(\frac{p(y,z,r|\theta_k)}{q_k(z)}\right)\mathrm{d}z,\ \forall q_k(z)\in\mathbb{S}\right\}.$$

The maximum value of

$$\int q_k(z)\log\left(\frac{p(y,z,r|\theta_k)}{q_k(z)}\right)\mathrm{d}z$$

over q_k is obtained when $q_k(z) = p(z|y, r, \theta) \in \mathbb{M}$, since the bound from above in (2) is attained for $p(z|y, r, \theta)$:

$$\int q(z) \log\left(\frac{p(y, z, r|\theta)}{q(z)}\right) \mathrm{d}z = \int p(z|y, r, \theta) \log\left(\frac{p(y, z, r|\theta)}{p(z|y, r, \theta)}\right) \mathrm{d}z = \log p(y, r|\theta).$$

Therefore, $\forall q_t(z) \in \mathbb{M}$,

$$\int q_t(z) \log\left(\frac{p(y, z, r|\theta_t)}{q_t(z)}\right) \mathrm{d}z \geq \max_{q_k \in \mathbb{M}^c} \int q_k(z) \log\left(\frac{p(y, z, r|\theta_k)}{q_k(z)}\right) \mathrm{d}z.$$

Now, suppose $\exists q_{w^*}(z) \in \mathbb{M}^c$ which maximizes Q over the set of wrong imputation models, and for which the equality holds in the above equation. This would imply that $q_{w^*}(z) \in \mathbb{M} \nsubseteq \mathbb{M}^c$ which would contradict with the assumption of $q_{w^*}(z) \in \mathbb{M}^c$.

Thus, $\forall q_t(z) \in \mathbb{M}$ and $\forall q_w(z) \in \mathbb{M}^c$,

$$\int q_t(z) \log\left(\frac{p(y, z, r|\theta_t)}{q_t(z)}\right) \mathrm{d}z > \int q_w(z) \log\left(\frac{p(y, z, r|\theta_w)}{q_w(z)}\right) \mathrm{d}z.$$

■

Lemma 2. *Let $q_0(z) \in \mathbb{M}^p$ be the most parsimonious correct imputation model and $q_1(z) \in \mathbb{M}/\mathbb{M}^p$ be an over-fitted correct imputation model. Partition*

$$\theta_1 = \begin{bmatrix} \theta_0 \\ \theta_r \end{bmatrix} = \begin{bmatrix} u \times 1 \\ s \times 1 \end{bmatrix}$$

and consider the hypothesis test $H_0 : \theta_r = \mathbf{0}$ versus $H_1 : \theta_r \neq \mathbf{0}$. Then, under H_0 and as $n \to \infty$,

$$2\left\{\sup_{\theta_1} \int q_1(z) \log\left(\frac{p(y, z, r|\theta_1)}{q_1(z)}\right) \mathrm{d}z - \sup_{\theta_0} \int q_0(z) \log\left(\frac{p(y, z, r|\theta_0)}{q_0(z)}\right) \mathrm{d}z\right\} \xrightarrow{d} \chi^2_s. \tag{A1}$$

Proof. Rewrite the null hypothesis as $H_0 : \theta_1 = g(\theta_0) = (\theta_0, \mathbf{0})^\intercal$ such that $G(\theta_0) = \partial g(\theta_0)/\partial \theta_0 = (I_u, \mathbf{0})^\intercal$ where I_u is the $u \times u$ identity matrix and $\mathbf{0}$ is a $s \times u$ matrix of zeros. Let $\hat{\theta}_1$ and $\hat{\theta}_0$ be the MLEs of θ_1 and θ_0, respectively. Also, let $S(\theta)$, $I(\theta)$ and $J(\theta)$ denote the full score function, Fisher and observed information matrix, respectively. Following Sen and Singer,[66] p. 205-207, if Assumptions 3 and 4 are satisfied and as $n \to \infty$, then a Taylor expansion around $\hat{\theta}_1$ yields

$$\begin{aligned} 2\int q_1(z) \log\left(\frac{p(y, z, r|\theta_1)}{q_1(z)}\right) \mathrm{d}z = 2\int p(z|y, r, \hat{\theta}_1) \log\left(\frac{p(y, z, r|\hat{\theta}_1)}{p(z|y, r, \hat{\theta}_1)}\right) \mathrm{d}z \\ - S^\intercal(\theta_1)[I(\theta_1)]^{-1} S(\theta_1) + o_p(1), \end{aligned}$$

since $[I(\theta_1)]^{-1} J(\hat{\theta}_1) \xrightarrow{p} I_{u+s}$ (Slutsky's theorem) where I_{u+s} is the identity matrix. Let $S^*(\theta)$, $I^*(\theta)$, and $J^*(\theta)$ denote the restricted score function, Fisher, and observed information matrix, respectively. Similarly, since $[I^*(\theta_0)]^{-1} J^*(\hat{\theta}_0) \xrightarrow{p} I_u$, we may write

$$\begin{aligned} 2\int q_0(z) \log\left(\frac{p(y, z, r|\theta_0)}{q_0(z)}\right) \mathrm{d}z = 2\int p(z|y, r, \hat{\theta}_0) \log\left(\frac{p(y, z, r|\hat{\theta}_0)}{p(z|y, r, \hat{\theta}_0)}\right) \mathrm{d}z \\ - S^{*\intercal}(\theta_0)[I^*(\theta_0)]^{-1} S^*(\theta_0) + o_p(1). \end{aligned}$$

Now, let $\mathrm{E}_{\theta_k}[.|\theta_k]$ be the expectation w.r.t. the correct imputation model $p(z|y, r, \theta_k)$. Since $\mathrm{E}_{\theta_1}\left[\log p(y, r|\theta_1)\right] - \mathrm{E}_{\theta_0}\left[\log p(y, r|\theta_0)\right] = 0$ under H_0, we have

$$\begin{aligned} &2\left\{\int p(z|y, r, \hat{\theta}_1) \log\left(\frac{p(y, z, r|\hat{\theta}_1)}{p(z|y, r, \hat{\theta}_1)}\right) \mathrm{d}z - \int p(z|y, r, \hat{\theta}_0) \log\left(\frac{p(y, z, r|\hat{\theta}_0)}{p(z|y, r, \hat{\theta}_0)}\right) \mathrm{d}z\right\} \\ &= S^\intercal(\theta_1)[I(\theta_1)]^{-1} S(\theta_1) - S^{*\intercal}(\theta_0)[I^*(\theta_0)]^{-1} S^*(\theta_0) + o_p(1). \end{aligned} \tag{A2}$$

Note that, under H_0,

$$S^*(\theta_0) = G^\intercal(\theta_0)S(\theta_1).$$

Also, since

$$S^*(\theta_0) \xrightarrow{d} N(0, I^*(\theta_0))$$

and

$$G^\intercal(\theta_0)S(\theta_1) \xrightarrow{d} N(0, G^\intercal(\theta_0)I(\theta_1)G(\theta_0)),$$

under H_0 we may write

$$I^*(\theta_0) = G^\intercal(\theta_0)I(\theta_1)G(\theta_0). \tag{A3}$$

It follows that (A2) may be simplified to

$$S^\intercal(\theta_1)\left[I^{-1}(\theta_1) - G(\theta_0)[I^*(\theta_0)]^{-1}G^\intercal(\theta_0)\right]S(\theta_1) + o_p(1). \tag{A4}$$

Furthermore,

$$\begin{aligned} tr\{I^{-1}(\theta_1) - G(\theta_0)[I^*(\theta_0)]^{-1}G^\intercal(\theta_0)\}I(\theta_1) &= tr\{I_{u+s} - G(\theta_0)[I^*(\theta_0)]^{-1}G^\intercal(\theta_0)I(\theta_1)\} \\ &= u + s - tr\{[I^*(\theta_0)]^{-1}G(\theta_0)I(\theta_1)G^\intercal(\theta_0)\} \\ &= u + s - tr\{[I^*(\theta_0)]^{-1}I^*(\theta_0)\} = s \end{aligned} \tag{A5}$$

Thus, using (A4) and (A5), and having $[I^*(\theta_0)]^{-1/2}S^*(\theta_0) \xrightarrow{d} N(0, I_u)$, and $[I(\theta_1)]^{-1/2}S(\theta_1) \xrightarrow{d} N(0, I_{s+u})$ by Slutsky's theorem, we obtain (A1).

■

Theorem 1. *Suppose M_0 is the imputation model chosen by BIC and $\mathbb{M}^p$ is a finite set of the most parsimonious correct models. If Assumptions 1 and 2 are satisfied, then*

$$\Pr(M_0 \in \mathbb{M}^p) \to 1 \quad as \quad n \to \infty.$$

Proof. Let $M_1 \in \mathbb{S}/\mathbb{M}^p$ be an arbitrarily chosen model which is not in the class of the most parsimonious models. To prove Theorem 1, it is sufficient to show that

$$\Pr(\text{BIC}(M_0) - \text{BIC}(M_1) < 0) \to 1 \; as \; n \to \infty. \tag{A6}$$

We have $\mathbb{S} = \mathbb{M} \bigcup \mathbb{M}^c$, where $\mathbb{M}^c$ is the complement of $\mathbb{M}$. Thus, either $M_1 \in \mathbb{M}/\mathbb{M}^p$ or $M_1 \in \mathbb{M}^c$. We show that (A6) holds under both of the following possible cases:

i. $M_1 \in \mathbb{M}/\mathbb{M}^p$
We need to show that the logarithmic penalty term in BIC outgrows the difference in the log-likelihood terms as $n \to \infty$. In this case, both M_0 and M_1 are correct, but M_1 is over-fitted w.r.t. to M_0, that is, $|\theta_1| - |\theta_0| > 0$.
Based on Lemma 2, we may write

$$\begin{aligned} &\lim_{n\to\infty} \Pr\left(\text{BIC}(M_0) - \text{BIC}(M_1) < 0\right) \\ &= \lim_{n\to\infty} \Pr\left(2\left\{\sup_{\theta_1}\int q_1(z)\log\left(\frac{p(y,z,r|\theta_1)}{q_1(z)}\right)dz - \sup_{\theta_0}\int q_0(z)\log\left(\frac{p(y,z,r|\theta_0)}{q_0(z)}\right)dz\right\}\right. \\ &\left. < \log(n)\times(|\theta_1| - |\theta_0|)\right) = \lim_{n\to\infty}\Pr\left(\chi^2_{|\theta_1|-|\theta_0|} < \log(n)\times(|\theta_1| - |\theta_0|)\right) = 1. \end{aligned}$$

ii. $M_1 \in \mathbb{M}^c$

Denote $KL(q_k||p)$ as the relative Kullback-Leibler divergence of model $q_k(z)$ from $p(z|y, r, \theta_k)$. In this case, M_0 is correct and M_1 is not, that is, $KL(q_1||p) > KL(q_0||p)$. Thus, we need to show that the difference in the log-likelihood terms in BIC outgrows the logarithmic penalty term as $n \to \infty$.

If Assumptions 1 and 2 are satisfied, then

$$\begin{aligned}
\frac{1}{2}\{\mathrm{BIC}(M_1) - \mathrm{BIC}(M_0)\} &= -\sup_{\theta_1}\left[(n-m)\left\{\frac{1}{n-m}\int q_1(z)\log\left(\frac{p(y,z,r|\theta_1)}{q_1(z)}\right)\mathrm{d}z\right\}\right. \\
&\quad \left. + m\left\{\frac{1}{m}\int q_1(z)\log\left(\frac{q_1(z)}{p(z|y,r,\theta_1)}\right)\mathrm{d}z\right\}\right] \\
&\quad + \sup_{\theta_0}\left[(n-m)\left\{\frac{1}{n-m}\int q_0(z)\log\left(\frac{p(y,z,r|\theta_0)}{q_0(z)}\right)\mathrm{d}z\right\}\right. \\
&\quad \left. + m\left\{\frac{1}{m}\int q_0(z)\log\left(\frac{q_0(z)}{p(z|y,r,\theta_0)}\right)\mathrm{d}z\right\}\right] \\
&\quad + \frac{1}{2}\log(n)\times(|\theta_1| - |\theta_0|) \\
&= (n-m)\left\{\frac{1}{n-m}\int q_0(z)\log\left(\frac{p(y,z,r|\theta_0)}{q_0(z)}\right)\mathrm{d}z\right. \\
&\quad \left. -\frac{1}{n-m}\int q_1(z)\log\left(\frac{p(y,z,r|\theta_1)}{q_1(z)}\right)\mathrm{d}z\right\} \\
&\quad - m\left\{\frac{1}{m}\int q_1(z)\log\left(\frac{q_1(z)}{p(z|y,r,\theta_1)}\right)\mathrm{d}z\right. \\
&\quad \left. -\frac{1}{m}\int q_0(z)\log\left(\frac{q_0(z)}{p(z|y,r,\theta_0)}\right)\mathrm{d}z\right\} + o_p(1) \\
&\quad + \frac{1}{2}\log(n)\times(|\theta_1| - |\theta_0|) \\
&= (n-m)\left\{\frac{1}{n-m}\int q_0(z)\log\left(\frac{p(y,z,r|\theta_0)}{q_0(z)}\right)\mathrm{d}z\right. \\
&\quad \left. -\frac{1}{n-m}\int q_1(z)\log\left(\frac{p(y,z,r|\theta_1)}{q_1(z)}\right)\mathrm{d}z\right\} \\
&\quad - m\left\{\frac{1}{m}KL(q_1||p) - \frac{1}{m}KL(q_0||p)\right\} + o_p(1) \\
&\quad + \frac{1}{2}\log(n)\times(|\theta_1| - |\theta_0|) \\
&= O_p(n-m) - O_p(m) \pm O(\log(\sqrt{n})) \\
&= O_p(n-m)
\end{aligned}$$

which tends to be positive since, according to Lemma 1 and the Law of Large Numbers, the term

$$(n-m)\left\{\frac{1}{n-m}\int q_0(z)\log\left(\frac{p(y,z,r|\theta_0)}{q_0(z)}\right)\mathrm{d}z - \frac{1}{n-m}\int q_1(z)\log\left(\frac{p(y,z,r|\theta_1)}{q_1(z)}\right)\mathrm{d}z\right\}$$

is positive and grows with probability approaching one as $n \to \infty$, with rate $O_p(n-m)$. Thus,

$$\Pr(\mathrm{BIC}(M_0) - \mathrm{BIC}(M_1) < 0) = \Pr(\mathrm{BIC}(M_1) - \mathrm{BIC}(M_0) > 0) \to 1 \ as \ n \to \infty.$$

■

Furthermore, to prove the following corollary, we require the following regularity condition:

Assumption 5. The initial values $\theta^{(0)}$ in the StEM algorithm is consistent asymptotically linear estimate of θ.

Corollary 1. *If the Assumptions 1 to 5 are satisfied, the BIC (see Section 3) computed at $\theta = \overline{\theta}$ is consistent for imputation model selection.*

Proof. Suppose $\hat{\theta}$ is the MLE of θ. It follows that

$$\sqrt{n}(\hat{\theta} - \theta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} [I(\theta)]^{-1} S_i(\theta) + o_p(1) \tag{A7}$$

as the sample size $n \to \infty$[66] p. 205-207.

Also, let $\overline{\theta}$ denote the StEM estimator of θ. Following Wang and Robins,[15] eq. A7 and the regularity condition in Assumption 5, we have

$$\sqrt{n}(\overline{\theta} - \theta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} [I(\theta)]^{-1} S_i(\theta) + o_p(1) \tag{A8}$$

as $n \to \infty$ and the number of imputations $M \to \infty$.

From (A7) and (A8), it follows that

$$\sqrt{n}(\overline{\theta} - \theta) = \sqrt{n}(\hat{\theta} - \theta) + o_p(1) \tag{A9}$$

as $n \to \infty$ and $M \to \infty$.

Hence, given the Assumptions 1 to 5 and from (A9), the results of Lemma 2, and consequently the result of Theorem 2, we may conclude that for sufficiently large number of imputations the BIC obtained from the StEM algorithm is consistent for imputation model selection. ∎

APPENDIX B. COMBINATION RULES

Below, we discuss combination rules for variance estimation of model parameters for MI and StEM. The purpose of this section is to compare the asymptotic results of their estimators and to provide a further heuristic argument to explain the connection between MI and StEM.

B.1 Rubin's rule

MI's combination rules (or Rubin's rule) are derived based on the Bayesian framework where θ and its estimate $\hat{\theta}$ are treated as unobserved random variables. Note that $\hat{\theta}$ denotes an estimate of θ in the absence of missing data. Further, with complete data, inferences about θ would be based on a normal approximation assumption $(\theta - \hat{\theta}) \sim N(0, W)$ where W is the associated variance of $(\theta - \hat{\theta})$. The mean and variance of the posterior distribution of θ are given by

$$\mathrm{E}(\theta|y, r) = \mathrm{E}(\hat{\theta}|y, r)$$

and

$$\mathrm{Var}(\theta|y, r) = \mathrm{E}(W|y, r) + \mathrm{Var}(\hat{\theta}|y, r), \tag{B1}$$

respectively. Based on this result, in the presence of missing data, the posterior mean and variance of θ can be approximated as described below.

Suppose that under a particular Bayesian model, $\hat{\theta}_1, \ldots, \hat{\theta}_j$ and $W_1, \ldots, W_j$ are the obtained values of $\hat{\theta}$ and W for each of $j = 1, \ldots, M$ imputed datasets which simulate features of posterior distribution of $\hat{\theta}$ and W. In addition, let $\overline{\theta}$ and B denote estimates of the posterior mean and variance of $\hat{\theta}$, respectively. Then, for an infinitely large number of imputations M, the combined estimate gives the MI estimator $\overline{\theta}$ which is the posterior mean of $\hat{\theta}$,

$$\overline{\theta} = \frac{1}{M} \sum_{j=1}^{M} \hat{\theta}_j = \mathrm{E}(\hat{\theta}|y, r).$$

The variability associated with this estimate is the sum of two components: the posterior variance of $\hat{\theta}$ and the posterior mean of W. The posterior variance of $\hat{\theta}$ is obtained by the between-imputation variance

$$B = \frac{1}{M-1}\sum_{j=1}^{M}(\hat{\theta}_j - \bar{\theta})^{\intercal}(\hat{\theta}_j - \bar{\theta}) = \mathrm{Var}(\hat{\theta}|y, r),$$

and the posterior mean of W is obtained by the within-imputation variance

$$W = \frac{1}{M}\sum_{j=1}^{M} W_j.$$

Thus, the trio statistic $(\bar{\theta}, \overline{W}, B)$ from the multiple imputed datasets provides the information we need to estimate the pair $(\hat{\theta}, W)$, and so to estimate θ.[24] Therefore, for a multiple imputation method to belong to the class of *proper* MI, multiple imputations must yield a consistent asymptotically normal estimator of θ and an unbiased estimator of its asymptotic variance when based on Rubin's rule. By using heuristic arguments and several examples, Rubin[8] concluded that "*Conclusion 4.1. If imputations are drawn to approximate repetitions from a Bayesian posterior distribution of missing data under the posited response* (missingness) *mechanism and an appropriate model for the data, then in large samples the imputation method is proper*", where the posterior distribution of missing data is defined as

$$p(z|y, r) = \int p(z|y, r, \theta)p(\theta|y, z)\mathrm{d}\theta. \tag{B2}$$

Hence, a non-Bayesian MI where its $\hat{\theta}_j$s are not random draws from their posterior distributions is an *improper* MI and its inference may not be based on Rubin's rule.

B.2 Louis' method

StEM's combination rules are carried out as follows. At each iteration, multiple imputations are randomly drawn from the imputation model given the current MLE of θ. Let $\hat{\theta}_1, \ldots, \hat{\theta}_j$ and $W_1, \ldots, W_j, j = 1, \ldots, M$, be the MLEs of θ and their variances, respectively, obtained from the next M iterations after the StEM algorithm converges. These final M imputations are implemented as multiple imputations in combination rules as shown in Diebolt and Ip[33] in a manner discussed below.

Let $\bar{\theta}$ be the StEM estimator which is the average of $\hat{\theta}_j$s over multiple imputed datasets, $\bar{\theta} = M^{-1}\sum_{j=1}^{M}\hat{\theta}_j$. Wang and Robins[15] and Nielsen[38] showed that, for sufficiently large M, $(\bar{\theta} - \theta) \sim N(0, I_{obs}^{-1})$ where I_{obs} denotes the Fisher information matrix. As such, the variance of the StEM estimator may be obtained based on the Louis' method,[67]

$$I_{obs} = \mathrm{E}\left[\frac{-\partial^2}{\partial\theta\partial\theta^{\intercal}}\log p(y, z, r|\theta)|y, r\right] - \mathrm{Var}\left[\frac{\partial}{\partial\theta}\log p(y, z, r|\theta)|y, r\right], \tag{B3}$$

and by replacing $\mathrm{E}[.\,|\,y, r]$ and $\mathrm{Var}[.\,|\,y, r]$ with their bootstrap estimates: from the difference between the complete information matrices of the $\hat{\theta}_j$s averaged over multiple imputed datasets *and* the variance of their respected score functions between multiple imputed datasets.

B.3 Connections between Rubin's rule and Louis' formula

Based on Rubin's rule, the MI variance estimator of $\bar{\theta}$ can be written as $\widehat{Var}(\theta) = W + B$ where as previously B is the between-imputation variance and W is the within-imputation variance. In EM terminology, W can be understood as a complete data estimate of the variance, related to the first term in Equation (B3), and B is closely related to the second term of (B3). Specifically, a StEM variance estimator of $\bar{\theta}$ was derived by von Hippel[68] which we can write as: $\widehat{Var}(\theta) = I_{obs}^{-1} = W(W - B)^{-1}W$.

APPENDIX C. SUPPLEMENTARY FOR SIMULATION STUDIES

C.1 Univariate missing predictor

The following plots and tables show the results of the simulation study in Section 4.1.1 for $\hat{\beta}_0$ and $\hat{\beta}_2$. Figure C1 and C2 show the magnitude of bias for $\hat{\beta}_0$ and $\hat{\beta}_2$, respectively. Table C1 and C2 show the RMSE for $\hat{\beta}_0$ and $\hat{\beta}_2$, respectively.

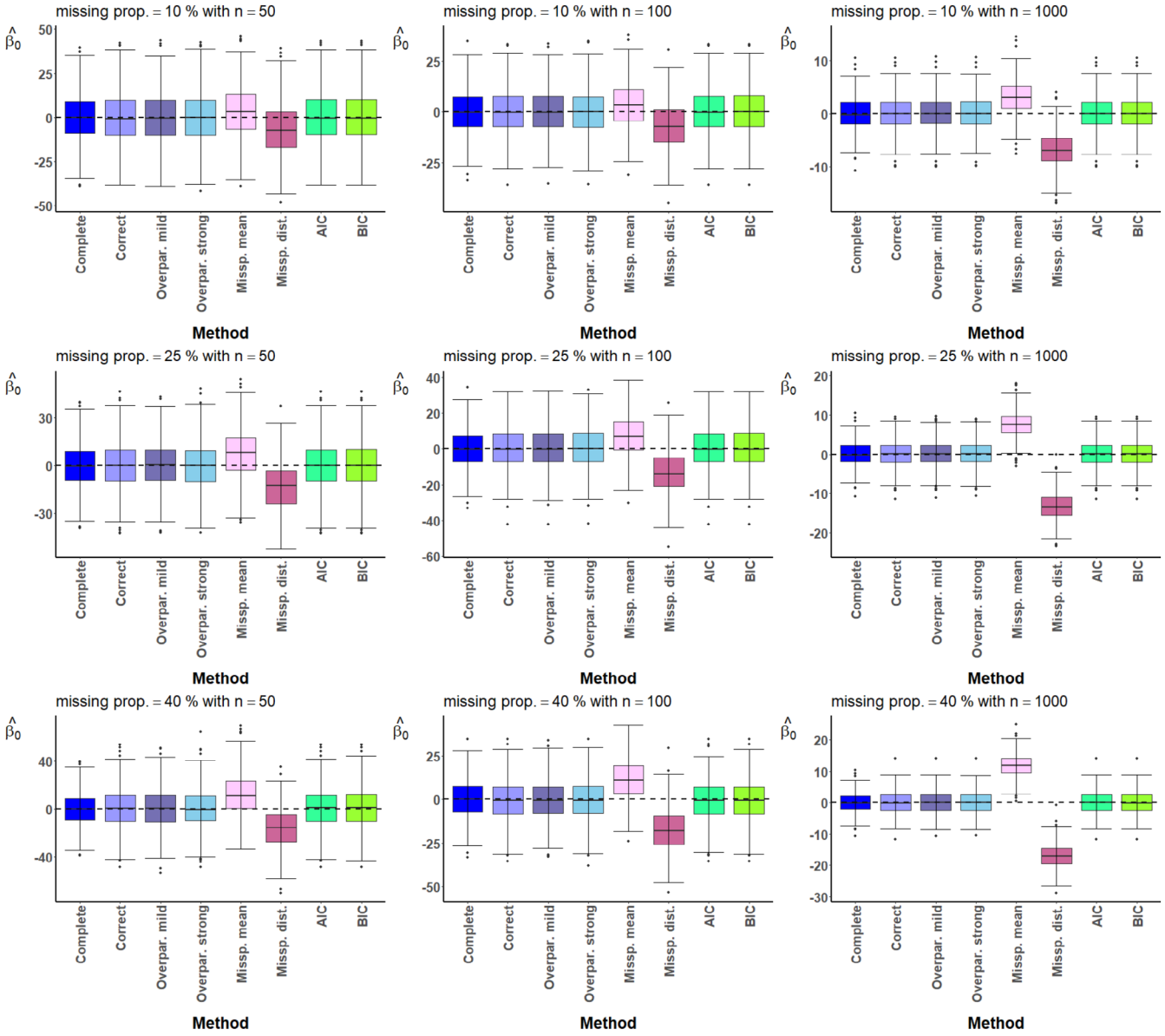

FIGURE C1 Boxplot of the relative bias (×100) of $\hat{\beta}_0$ for different methods. Each method is compared with the original dataset (Complete) based on the accuracy of their estimators as the missing proportion and the sample size increase from the top-left corner to the bottom-right. Note that the true value of the intercept coefficient is 1 [Colour figure can be viewed at wileyonlinelibrary.com]

Furthermore, the results of the simulation study in Section 4.1.1 for Amelia and MICE are presented in Figure C3 to C5 and Table C3 to C5.

C.2 Univariate missing response

We considered a log-linear Poisson regression model,

$$Y_i|X_{1i}, X_{2i} \overset{i.i.d}{\sim} \text{Poisson}(\lambda_i), \quad i = 1, \ldots, n,$$

with

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{1i} X_{2i}.$$

Here, in addition to having missingness in the response variable, we have added more complexity to the design by considering an interaction term in the model instead of an additive structure (c.f. simulation scenario (a) in Section 4.1.1). The

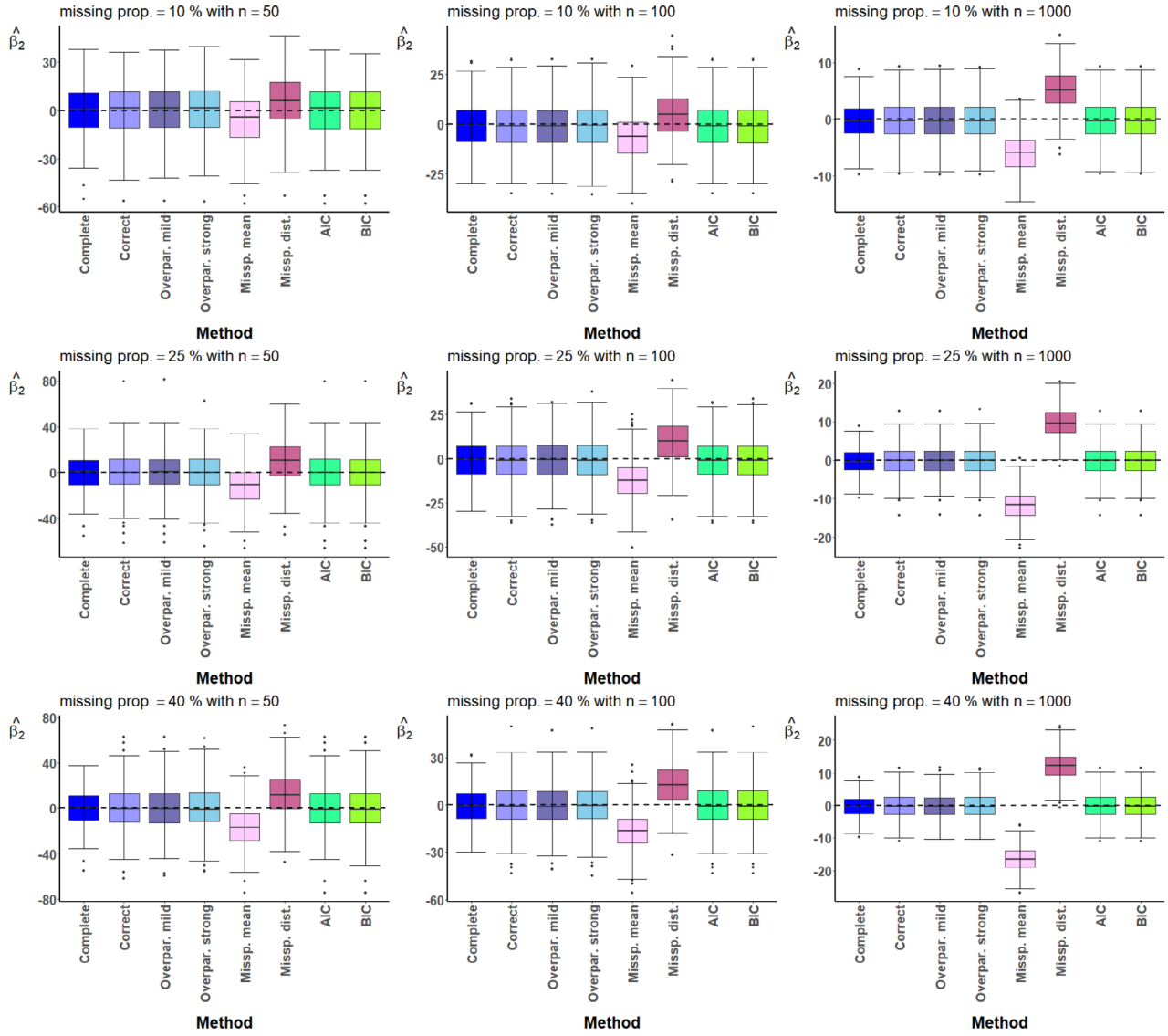

FIGURE C2 Boxplot of the relative bias (×100) of $\hat{\beta}_2$ for different methods. Each method is compared with the original dataset (Complete) based on the accuracy of their estimators as the missing proportion and the sample size increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1 [Colour figure can be viewed at wileyonlinelibrary.com]

predictors X_{1i} and X_{2i} are fully observed where $X_{1i} \sim$ Bernoulli(0.35) and $X_{2i} \sim$ Bernoulli(0.5) but the response is subject to missingness such that r_i is the missingness indicator variable of Y_i—for example, $r_i = 0$ if Y_i is missing and $r_i = 1$ if Y_i is observed. Let ϕ be the probabilities of $r_i = 0$, which satisfies $\text{logit}(\phi) = \phi_0 + \phi_1 X_{1i} + \phi_2 X_{2i} + \phi_3 X_{1i} X_{2i}$.

We ran 500 simulations and set $\beta = (1, 1, 1, 1)^\intercal$. We imposed 24% missingness, on average, in the response variable Y_i by setting $\phi = (-1, -0.1, -0.1, -0.5)^\intercal$. Once again, we used the squared distance of the parameters between the $(t+1)$th and tth iterations as the convergence criterion and set the convergence threshold to 10^{-4} with the number of multiple imputations set to $M = 100$.

To investigate the performance of information criteria for imputation model selection in Section C0.2, missing values $Y_{mis,i}$ were imputed under the following three candidate models and under MAR assumption:

Model 1: $Y_{mis,i} \sim \text{Poisson}(\lambda_i)$ where $\lambda_i = \beta_0 + \beta_1 X_{1i}$,
Model 2: $Y_{mis,i} \sim \text{Poisson}(\lambda_i)$ where $\lambda_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$,
Model 3: $Y_{mis,i} \sim \text{Poisson}(\lambda_i)$ where $\lambda_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{1i} X_{2i}$,
Model 4: $Y_{mis,i} \sim \text{Poisson}(\lambda_i)$ where $\lambda_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{1i} X_{2i} + \beta_3 X_{3i}$,

TABLE C1 RMSE (×1000) of $\hat{\beta}_0$ for different imputation methods across 500 simulations, compared with the original dataset (Complete)

	10%			25%			40%		
	n=50	n=100	n=1000	n=50	n=100	n=1000	n=50	n=100	n=1000
Complete	137	103	31	137	103	31	137	103	31
Correct	143	106	32	150	113	33	167	120	36
Overpar. mild	143	106	32	151	112	33	167	120	36
Overpar. strong	144	106	32	152	113	33	167	119	35
Missp. mean	149	112	46	170	136	82	206	162	124
Missp. dist.	162	130	75	202	174	137	232	213	174
AIC	143	106	32	150	113	33	168	120	36
BIC	143	106	32	151	113	33	168	120	36

Note: The true value of the slope coefficient is 1.

TABLE C2 RMSE (×1000) of $\hat{\beta}_2$ for different imputation methods across 500 simulations, compared with the original dataset (Complete)

	10%			25%			40%		
	n=50	n=100	n=1000	n=50	n=100	n=1000	n=50	n=100	n=1000
Complete	147	110	32	147	110	32	147	110	32
Correct	157	119	34	170	124	36	186	134	38
Overpar. mild	157	118	34	170	124	36	185	133	38
Overpar. strong	158	119	34	166	123	36	185	133	38
Missp. mean	169	131	68	197	169	122	236	203	168
Missp. dist.	170	132	64	204	161	105	226	187	129
AIC	157	119	34	173	124	36	190	133	38
BIC	157	119	34	173	125	36	190	134	38

Note: The true value of the slope coefficient is 1.

where X_{3i} is an auxiliary variable with $X_{3i} \sim \text{Gamma}(2, 2)$. In other words, we have assumed that Model 3 is correct, but Model 1 and Model 2 underfit the data because of underparameterization, that is, only one predictor X_{1i} is included (X_{2i} is ignored) in Model 1, and in Model 2, the interaction term $X_{1i}X_{2i}$ is ignored. Furthermore, model 4 overfits the data because of overparametrization by including X_{3i} in the model.

Results in Table C6 show the proportion of times the corresponding model is chosen based on each information criterion for various sample sizes of $n = \{50, 100, 1000\}$. Similarly to the simulation study with a univariate missing predictor, we see that for even a small sample of 50, both AIC and BIC were able to choose the correct model (Model 3) here at least 81% of the time. The performances of BIC improved as the sample size increased where it was able to choose the correct model at least 97% of the time for a sample size of $n = 100$ and larger.

Finally, we investigated the impact of imputation model selection on the post-selection inference from data. Table C7 shows the sample average of estimates of β_j, denoted as $\overline{\beta}_j, j = 0, 1, 2, 12$, for the corresponding imputation model averaged over 500 simulations *and* its mean square error, $\text{MSE}_j = (\overline{\beta}_j - \beta_j)^2 + s_j^2$, where s_j is the simulated standard error of $\overline{\beta}_j$. This table shows that more accurate estimates and smaller mean squared errors are obtained when using the correct imputation model (Model 3) as well as the over-fitted model (model 4). In many practical situations, it is recommended to add as many variables as possible to the imputation model when the imputation model is not known. In Table C6, we noted that AIC chooses the over-fitted model (Model 4) 14.8-16% of the time, irrespective of sample size. Results in

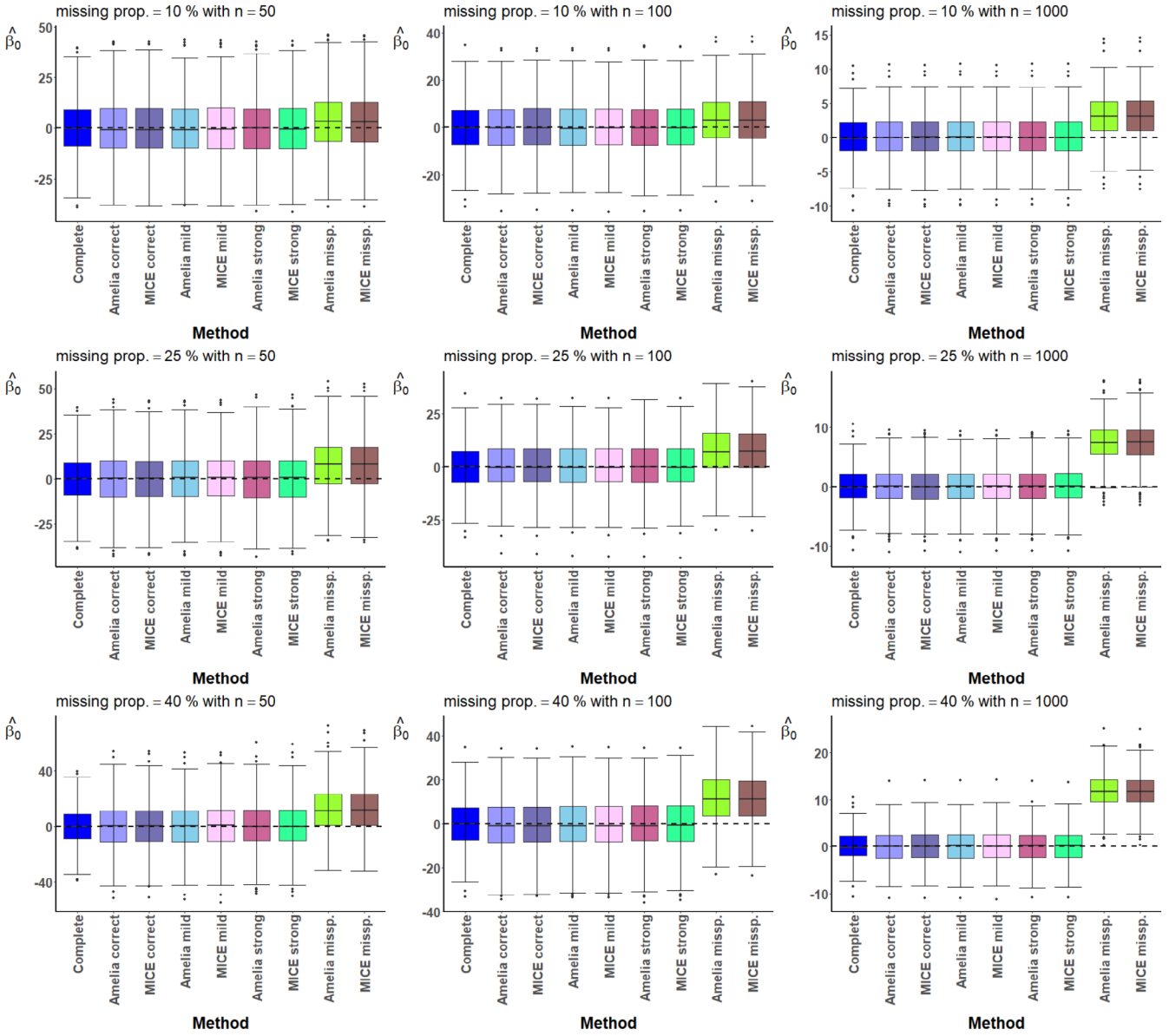

FIGURE C3 Boxplot of the relative bias (×100) of $\hat{\beta}_0$ for multiple imputation softwares. Amelia and MICE correct, overparametrized mild/strong and misspecified mean models are compared with the original dataset (Complete) based on the accuracy of their estimators as the missing proportion and the sample size increase from the top-left corner to the bottom-right. Note that the true value of the intercept coefficient is 1 [Colour figure can be viewed at wileyonlinelibrary.com]

Table C7 suggest that perhaps overfitting has little impact on post-selection inference. However, the reader concerned about the post-selection impact of overfitting when using AIC can use an information criterion with consistency property such as BIC (see Theorem 1).

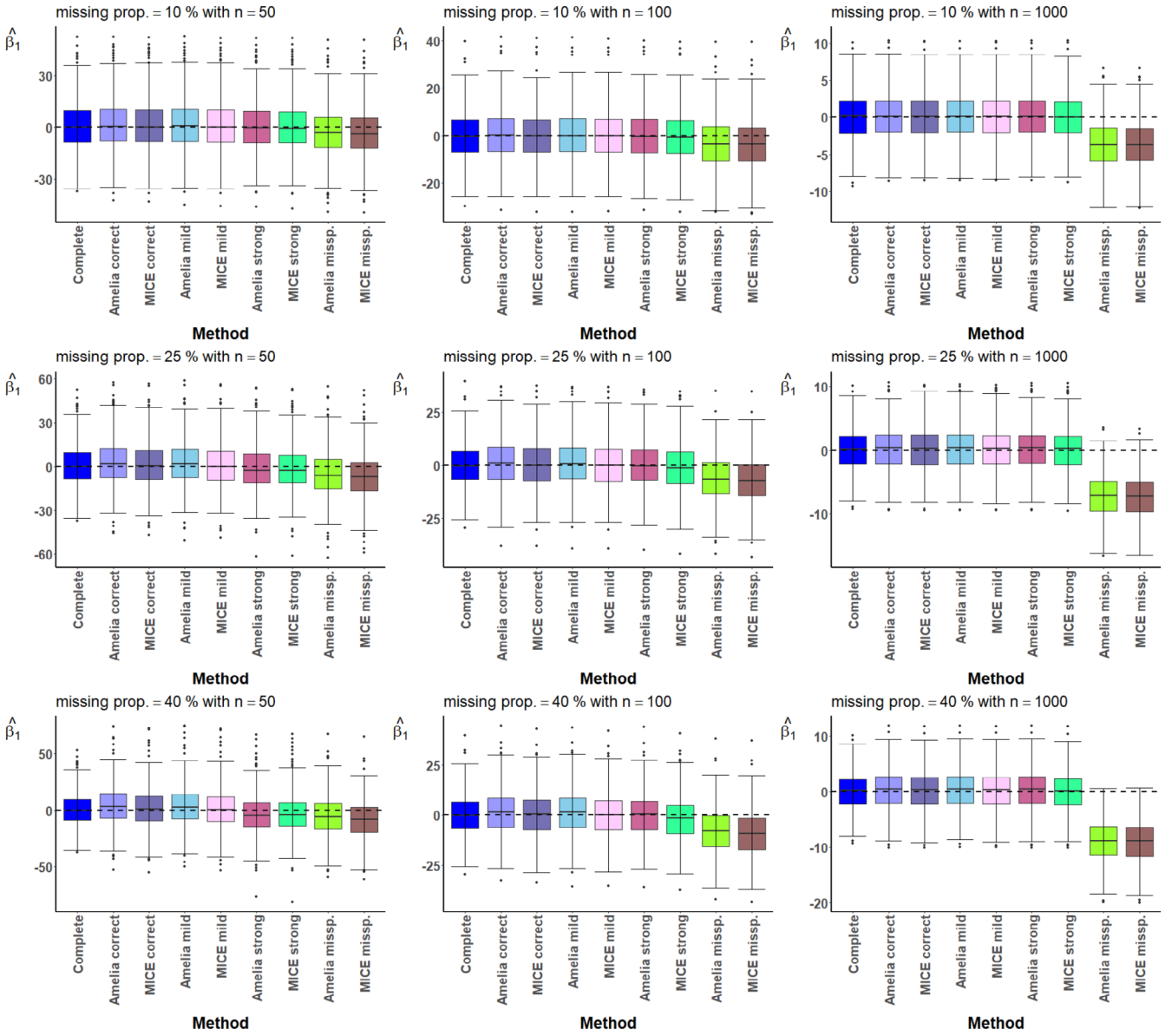

FIGURE C4 Boxplot of the relative bias (×100) of $\hat{\beta}_1$ for multiple imputation softwares. Amelia and MICE correct, overparametrized mild/strong and misspecified mean models are compared with the original dataset (Complete) based on the accuracy of their estimators as the missing proportion and the sample size increase from the top-left corner to the bottom-right. Note that the true value of the intercept coefficient is 1 [Colour figure can be viewed at wileyonlinelibrary.com]

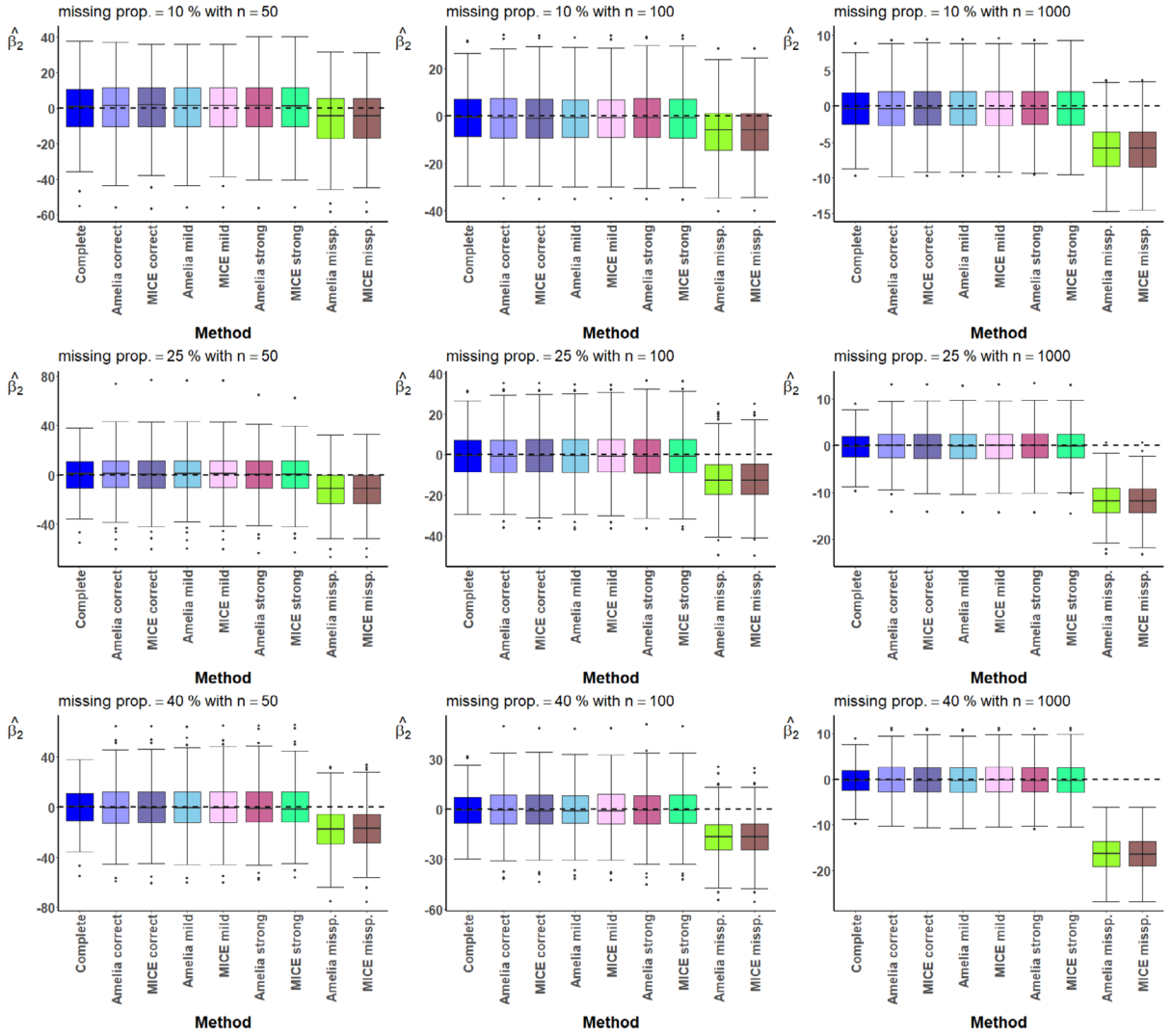

FIGURE C5 Boxplot of the relative bias (×100) of $\hat{\beta}_2$ for multiple imputation softwares. Amelia and MICE correct, overparametrized mild/strong and misspecified mean models are compared with the original dataset (Complete) based on the accuracy of their estimators as the missing proportion and the sample size increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE C3 RMSE (×1000) of $\hat{\beta}_0$'s for Amelia and MICE correct, overparametrized mild/strong and misspecified mean models across 500 simulations, compared with the original dataset (Complete)

	10%			25%			40%		
	n=50	*n*=100	*n*=1000	*n*=50	*n*=100	*n*=1000	*n*=50	*n*=100	*n*=1000
Complete	137	103	31	137	103	31	137	103	31
Amelia correct	144	106	32	151	113	33	168	121	36
MICE correct	143	106	32	149	112	33	166	120	36
Amelia mild	143	106	32	150	112	33	167	120	35
MICE mild	143	106	32	150	112	33	167	119	36
Amelia strong	144	107	32	152	113	33	167	121	36
MICE strong	145	107	32	152	113	33	167	120	35
Amelia missp.	149	112	45	171	137	82	207	164	124
MICE missp.	149	112	45	170	136	82	205	162	124

Note: The true value of the coefficients is 1.

TABLE C4 RMSE (×1000) of $\hat{\beta}_1$'s for Amelia and MICE correct, overparametrized mild/strong and misspecified mean models across 500 simulations, compared with the original dataset (Complete)

	10%			25%			40%		
	***n*=50**	***n*=100**	***n*=1000**	***n*=50**	***n*=100**	***n*=1000**	***n*=50**	***n*=100**	***n*=1000**
Complete	146	103	32	146	103	32	146	103	32
Amelia correct	153	107	32	161	111	34	177	117	36
MICE correct	153	106	32	159	111	34	172	115	36
Amelia mild	153	106	32	161	111	34	176	116	36
MICE mild	153	107	32	159	111	34	172	115	36
Amelia strong	152	107	32	161	111	34	181	116	36
MICE strong	152	107	32	161	111	34	179	117	36
Amelia missp.	152	113	50	165	129	80	182	142	96
MICE missp.	152	114	50	170	133	81	189	149	98

Note: The true value of the coefficients is 1.

TABLE C5 RMSE (×1000) of $\hat{\beta}_2$'s for Amelia and MICE correct, overparametrized mild/strong and misspecified mean models across 500 simulations, compared with the original dataset (Complete)

	10%			25%			40%		
	***n*=50**	***n*=100**	***n*=1000**	***n*=50**	***n*=100**	***n*=1000**	***n*=50**	***n*=100**	***n*=1000**
Complete	147	110	32	147	110	32	147	110	32
Amelia correct	157	119	34	170	124	36	188	134	38
MICE correct	157	119	34	170	124	36	185	133	38
Amelia mild	157	119	34	169	124	36	188	133	38
MICE mild	157	119	34	170	124	36	186	133	38
Amelia strong	158	119	34	167	124	36	187	135	38
MICE strong	158	119	34	167	123	36	189	134	38
Amelia missp.	169	131	68	198	170	123	241	206	168
MICE missp.	169	131	68	197	170	123	238	205	168

Note: The true value of the coefficients is 1.

TABLE C6 Proportion of times (%) the information criterion chooses the fitted imputation model for different sample sizes in 500 simulated datasets

		Model 1	Model 2	Model 3	Model 4
n=50	AIC	0.0	3.0	**81.0**	16.0
	BIC	0.0	10.0	**83.8**	6.2
n=100	AIC	0.0	0.0	**85.2**	14.8
	BIC	0.0	0.2	**97.0**	2.8
n=1000	AIC	0.0	0.0	**84.4**	15.6
	BIC	0.0	0.0	**99.0**	1.0

Note: Model 3 (in bold) is the correct model, which was selected with the highest proportion even for a small sample size of $n = 50$. Note that the proportion of times this model was chosen approached one as sample size increased.

TABLE C7 $\overline{\beta}_j$ and mean square error (in parenthesis) for different imputation models across 500 simulations for different sample sizes

		$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_{12} = 1$
$n = 50$	Model 1	1.171 (0.09)	1.423 (0.47)	0.719 (0.13)	0.627 (0.42)
	Model 2	0.819 (0.10)	1.284 (0.18)	1.187 (0.11)	0.713 (0.19)
	Model 3	0.938 (0.07)	1.095 (0.14)	1.033 (0.09)	0.945 (0.16)
	Model 4	0.937 (0.07)	1.095 (0.14)	1.034 (0.09)	0.945 (0.16)
$n = 100$	Model 1	1.193 (0.07)	1.438 (0.33)	0.708 (0.11)	0.601 (0.29)
	Model 2	0.862 (0.05)	1.228 (0.10)	1.158 (0.06)	0.754 (0.11)
	Model 3	0.982 (0.04)	1.029 (0.06)	0.999 (0.05)	0.998 (0.07)
	Model 4	0.981 (0.04)	1.030 (0.06)	0.999 (0.04)	0.997 (0.07)
$n = 1000$	Model 1	1.211 (0.05)	1.397 (0.17)	0.709 (0.09)	0.618 (0.16)
	Model 2	0.878 (0.02)	1.192 (0.04)	1.161 (0.03)	0.764 (0.06)
	Model 3	1.001 (0.00)	0.999 (0.01)	1.000 (0.00)	1.000 (0.01)
	Model 4	1.001 (0.00)	0.999 (0.01)	1.000 (0.00)	1.000 (0.01)